EXHIBIT 1

GNE.2930R1C1

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant	:	Botstein, et al.
Appl. No.	:	9/866,034
Filed	:	May 25, 2001
For	:	SECRETED AND TRANSMEMBRANE POLYPEPTIDES AND NUCLEIC ACIDS ENCODING THE SAME
Examiner	:	Spector, L.
Group Art Unit	:	1647

DECLARATION OF VICTORIA SMITH, Ph.D., UNDER 37 CFR §1.132

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

I, Dr. Victoria Smith, declare and state as follows:

1. I am a Senior Scientist in the Department of Molecular Biology of Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080.

2. My scientific Curriculum Vitae, including my list of publications, is attached to and forms part of this Declaration (Exhibit A).

3. I joined Genentech in 1996. For approximately three years, I directed a laboratory in the Department of Molecular Biology. During this time I was involved in target discovery for the Tumor Antigen Project, using DNA microarrays to discover genes differentially expressed in tumors compared to their expression in normal tissues. In connection with the above-identified patent application, I directed the generation and analysis of the microarray data attached as Exhibit B.

4. Exhibit B reports the results of the microarray analysis conducted on the gene encoding PRO1800 (DNA35672) as part of the investigation of several newly discovered DNA sequences. The column "Unq Id" identify the gene as 851, which is DNA35672, while the column "DNA Id" identifies the particular lot of PCR product used. The microarray experiments were performed using well-established and accepted microarray techniques known in the art. (*See, e.g.*, Nature Revs. Genetics, 5:229-237 (2004), attached as Exhibit C). The DNA samples used in the microarray studies were obtained from individual lung tumor tissue samples or individual normal lung tissue samples. The individual tumor and normal lung samples were each

compared to pooled samples of normal epithelial tissue. The level of expression in the lung tumor or normal lung tissue sample was compared to the normal pooled epithelial sample, and reported as a raw ratio. The average of the normal lung samples was then used to normalize the data to generate a ratio of expression of the PRO1800 gene in lung tumor samples compared to the average expression in normal lung tissue. In the results reported in Exhibit B, a ratio of 2.0 or greater is a significant result, and indicates a significant increase in expression of the PRO1800 gene in lung tumor tissue compared to the normal lung tissue controls.

5. The results of the microarray studies reported in Exhibit B indicate that the gene encoding PRO1800 (DNA35672) is significantly overexpressed in nine of the eighty lung tumor samples tested compared to the normal lung tissue controls. That is the equivalent of one in every nine samples. In contrast, none of the individual normal lung tissue samples show significant overexpression of the PRO1800 gene. In addition, the average ratio of the lung tumor samples is significantly different from the average ratio of the individual normal lung tumor samples ($p < 0.01$).

6. It is well-established in the art that overexpression of the mRNA for a gene is likely to lead to overexpression of the corresponding protein. Support for this statement can be found, for example, in the Molecular Biology of the Cell, a leading textbook in the field. (Bruce Alberts, *et al.*, Molecular Biology of the Cell (4th ed. 2002), excerpts submitted herewith as Exhibit D). Figure 6-3 on page 302 illustrates the basic principle that there is a correlation between increased gene expression and increased protein expression. The accompanying text states that "a cell can change (or regulate) the expression of each of its genes according to the needs of the moment – *most obviously by controlling the production of its mRNA*." Molecular Biology of the Cell at 302, emphasis added. Similarly, figure 6-90 on page 364 illustrates the path from gene to protein. The accompanying text states that while potentially each step can be regulated by the cell, "the initiation of transcription is the most common point for a cell to regulate the expression of each of its genes." Molecular Biology of the Cell at 364. This point is repeated on page 379, where the authors state that of all the possible points for regulating protein expression, "[f]or most genes transcriptional controls are paramount." Molecular Biology of the Cell at 379.

7. While not every lung tumor sample tested shows overexpression of the PRO1800 gene, the data in Exhibit B indicate that a significant portion of lung tumors do (one in every nine), while none of the normal lung tissue samples show overexpression. Given the known correlation between overexpression of a gene and the corresponding overexpression of the encoded protein, it is very likely that a similar number of lung tumors will overexpress the PRO1800 protein, while normal lung tissue samples will not. Together with the data reported in Example 16 that the gene encoding PRO1800 is amplified in some lung tumors, the results reported in Exhibit B indicate that the PRO1800 gene and protein, as well as antibodies to the encoded protein, can be used to differentiate some cancerous lung tissue from normal lung tissue. Because not all lung tumors show overexpression of PRO1800, it cannot be used to exclude a sample being tested as non-cancerous. However, the PRO1800 gene, protein, and corresponding antibodies are useful as a diagnostic tool since a significant portion of lung tumors overexpress the gene and most likely the encoded protein, while no normal lung samples do.

8. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information or belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful statements may jeopardize the validity of the application or any patent issued thereon.

By: _____ Date: 01/20/05
Victoria Smith, Ph.D.

S:\DOCS\BSG\BSG-1773.DOC
011905

VICTORIA SMITH

Genentech Inc.
Dept. Molecular Biology
1 DNA Way
South San Francisco CA 94080
Ph: (650) 225 7382
Fax: (650) 225 6497
Email: victoria@gene.com

EDUCATION

Ph.D. (1991) Molecular Biology, University of Cambridge, Cambridge, United Kingdom.
Honors (1987) Biochemistry, University of Western Australia, Australia.
Bachelor of Science (1986) Physical and Inorganic Chemistry, and Biochemistry, University of Western Australia, Australia.

WORK AND RESEARCH EXPERIENCE

Senior Scientist, Genentech Inc (August 1996 – present; promoted to Senior Scientist March 2001)
Lab head, Dept. Molecular Biology
- Identification of potential therapeutic targets for cancer using novel microarray technology
- Discovery and identification of novel secreted proteins
- Development of cancer therapuetics

Stanford University, California, U.S.A. (February 1995 – August 1996)
Research Fellow, Department of Biochemistry
Research: Genomic functional analysis of chromosome V of *Saccharomyces cerevisiae*

Stanford University, California, U.S.A. (January 1992 - January 1995)
Postdoctoral Fellow, Department of Genetics.

Research: Development of new methodology for whole genome functional analysis in microorganisms using genomic sequence data and insertional mutagenesis.

Awards

Human Frontiers Science Program Organization Long Term Fellowship (accepted, 4/01/93 - 1/31/95).

American Cancer Society (California Division) Fellowship (1993, declined).

Cambridge University, United Kingdom (October 1988 - December 1991)

Research undertaken at the Medical Research Council Laboratory of Molecular Biology, Cambridge, UK, for the degree of Doctor of Philosophy, Cambridge University.

Thesis: A Molecular Genetic Analysis of Yeast Chromosome IX.

Thesis Advisor: Dr. Barclay Barrell.

Awards

Max Perutz Prize in 1991 for outstanding performance by a graduate student. Awarded for advances in genomic-scale DNA sequencing methodology, and genetic analysis of the *SNP1* gene of *Saccharomyces cerevisiae* .

King Edward Memorial Hospital for Women, Western Australia (1988)

Research Scientist.

Research: Analysis of hormone-inducible mRNAs in breast tumors.

University of Western Australia (1984 - 1987)

1984 - 1986: Bachelor of Science degree with double major in Biochemistry and Physical and Inorganic Chemistry.

1987: First Class Honors in Biochemistry. Thesis: Nuclease Sensitivity and Methylation Patterns of the Phenylalanine Hydroxylase Gene.

Awards

Association of Commonwealth Universities Scholarship for study in the United Kingdom (accepted, October 1988 - October 1991, University of Cambridge).

Hackett Scholarship for overseas study (1988, declined).

Lady James Prize in Natural Science in 1986: Best student completing Bachelor of Science degree with a major in a natural science.

JWH Lugg Prize in Biochemistry in 1986: Best student completing major in Biochemistry.

Convocation Prize in Science in 1985: Best student completing major in second year Physics, Geology or Chemistry.
Shell Prize in Chemistry in 1985: Best student completing major in second year Chemistry.

PUBLICATIONS

V. Smith, E.F. Shen, D. Wieand, T.H. Landon, N.A. Wong, A.M. Lessells, S. Paterson-Brown, T.D. Wu, J.Z. Tang, K.J. Hillan, I.D. Penman, Expression Analysis of the Metaplasia-Dysplasia-Carcinoma Sequence in Barrett's Esophagus (submitted).

Tice DA. Szeto W. Soloviev I. Rubinfeld B. Fong SE. Dugger DL. Winer J. Williams PM. Wieand D. Smith V. Schwall RH. Pennica D. Polakis P. *Journal of Biological Chemistry.* 277(16):14329-35, 2002 Apr 19.

N.J. Maughan, F. Lewis, V. Smith (2001) *Journal of Pathology* 195, 3-6.

F. Lewis, N.J. Maughan, V. Smith, K. Hillan, P. Quirke (2001) *Journal of Pathology* 195, 66-71.

D.J. Garfinkel, M.J. Curcio and V. Smith (1998) Ty Mutagenesis *Methods in Microbiology,* volume 26, 101-117.

C. Churcher, S. Bowman, K. Badcock, A. Bankier, D. Brown, T. Chillingworth, R. Connor, K. Devlin, S. Gentles, N. Hamlin, D. Harris, T. Horsnell, S. Hunt, K. Jagels, M. Jones, G. Lye, S. Moule, C. Odell, D. Pearson, M. Rajandream, P. Rice, N. Rowley, J. Skelton, V. Smith, S. Walsh, S. Whitehead & B. Barrell. (1997) *Nature,* 387, 84-87.

F. S. Dietrich, J. Mulligan, K. Hennessy, M. A. Yelton, E. Allen, R. Araujo, E. Aviles, A. Berno, T. Brennan, J. Carpenter, E. Chen, J. M. Cherry, E. Chung, M. Duncan, E. Guzman, G. Hartzell, S. Hunicke-Smith, R. W. Hyman, A. Kayser, C. Komp, D. Lashkari, H. Lew, D. Lin, D. Mosedale, K. Nakahara, A. Namath, R. Norgren, P. Oefner, C. Oh, F.X. Petel, D. Roberts, P. Sehl, S. Schramm, T. Shogren, V. Smith, P. Taylor, Y. Wei, D. Botstein & R. W. Davis. (1997) *Nature* 387, 78-81.

V. Smith, K. Chou, D. Lashkari, D. Botstein, and P. O. Brown. (1996). Functional Analysis of the Genes of Yeast Chromosome V by Genetic Footprinting. *Science* 274, 2069-74

V. Smith, D. Botstein, and P. O. Brown (1995) Genetic Footprinting: A genomic strategy for determining a gene's function given its sequence. *Proc. Natl. Acad. Sci. U.S.A.*, 92, 6479-6483.

V. Smith, M. Craxton, A. T. Bankier, C. M. Brown, W. D. Rawlinson, M. Chee, and B. G. Barrell (1995) Microtiter methods for the preparation and fluorescent sequencing of M13 clones. *Recombinant DNA Methodology II: a volume in the Selected Methods in Enzymology Series,* pp. 607 - 621.

V. Smith, M. Craxton, A. T. Bankier, C. M. Brown, W. D. Rawlinson, M. S. Chee, and B. G. Barrell (1993) Microtiter methods for the preparation and fluorescent sequencing of M13 clones. *Methods in Enzymology,* 218 , 173-187.

V. Smith and M. S. Chee (1991) A simple method for sequencing the complementary strand of ssDNA from M13 clones. *Nucleic Acids Research* 19, 6957.

V. Smith and B. G. Barrell (1991) Cloning of a Yeast U1 snRNP 70K protein homologue: functional conservation of an RNA binding domain between humans and yeast. *EMBO Journal* 10, 2627-2643.

W. D. Rawlinson, M. S. Chee, V. Smith and B. G. Barrell (1991) Preparation of large numbers of single stranded DNA templates by rescue from phagemids in microtiter plates. *Nucleic Acids Research* 19, 4779.

V. Smith, C. M. Brown, A. T. Bankier, and B. G. Barrell (1990) Semi-automated preparation of DNA templates for large scale sequencing projects. *DNA Sequence* 1, 73-78.

S. J. Wysocki, E. Hahnel, S. P. Wilkinson, V. Smith, and R. Hahnel (1990) Hormone-sensitive gene expression in breast tumors. *Anticancer Research* 10, 185-188

S. J. Wysocki, E. Hahnel, A. Masters, V. Smith, A. J. McCartney and R. Hahnel (1990) Detection of pS2 messenger RNA in gynecological cancers. *Cancer Research* 50, 1800-1802

PATENTS

Genetic Footprinting: Insertional Mutagenesis and Genetic Selection. U.S. Patent No. 5,612,180. Inventors: Patrick Brown and Victoria Smith

PATENTS FILED (at Genentech Inc.)

Methods of Detecting and Quantifying Gene Expression. Inventors: Victoria Smith, Edward Robbie, David Lowe, James Marsters.

Compositions and Methods for the Treatment of Cancer. Inventors: Victoria Smith, Austin Gurney, Audrey Goddard, Fred DeSauvage

Diagnostic for Dysplasia in Barrett's Esophagus. Inventor: Victoria Smith

Numerous composition of matter filings related to novel gene discovery, pending

Unq Id	DNA Id	Experiment Name	Raw Ratio (sample/pooled epithelial)	Normalized Ratio (sample/normal lung)
Lung Tumor Samples				
851.	157,736.	100ngLungBaCa1069 vs 25ngEpi1409	0.563	0.990
851.	157,736.	lung 1431 vs epi pool	0.634	1.115
851.	157,736.	lung SqCa R694 vs epi pool	0.847	1.489
851.	157,736.	Lung SqCa-hf 1649	0.337	0.593
851.	157,736.	Lung SqCa-hf 1649	0.209	0.367
851.	157,736.	lung tumor 1055 vs epi pool	1.085	1.908
851.	157,736.	lung tumor 10ng	0.858	1.509
851.	157,736.	lung tumor 1370 vs epi pool	0.404	0.710
851.	135,920.	lung tumor 1647	0.444	0.781
851.	135,920.	lung tumor 1648	0.633	1.113
851.	157,736.	lung tumor 1685/ref.RNA	0.886	1.558
851.	135,920.	Lung Tumor 685	0.18	0.316
851.	157,736.	lung tumor 688 vs epi pool	0.443	0.779
851.	135,920.	lung tumor 734	0.456	0.802
851.	135,920.	lung tumor 735	2.136	3.756
851.	135,920.	lung tumor 737	0.789	1.387
851.	135,920.	lung tumor 738	0.884	1.554
851.	135,920.	lung tumor 739	0.655	1.152
851.	101,241.	Lung tumor hf 842	0.849	1.493
851.	135,920.	Lung Tumor HF-0017083	0.467	0.821
851.	135,920.	lung tumor hf-1291	0.078	0.137
851.	157,736.	lung tumor hf-1333	0.496	0.872
851.	157,736.	lung tumor hf-1340	0.514	0.904
851.	157,736.	lung tumor hf-1364	0.343	0.603
851.	157,736.	lung tumor hf-1366	0.634	1.115
851.	157,736.	lung tumor hf-1587	0.478	0.840
851.	157,736.	lung tumor hf-1587 v normal	0.442	0.777
851.	157,736.	lung tumor hf-1587 v. normal	0.404	0.710
851.	135,920.	lung tumor hf-1596	0.491	0.863
851.	135,920.	lung tumor hf-1646	0.258	0.454
851.	135,920.	lung tumor hf-1649	0.278	0.489
851.	135,920.	lung tumor hf-1655	0.597	1.050
851.	135,920.	lung tumor hf-1719	0.612	1.076
851.	157,736.	lung tumor hf-1775	0.422	0.742
851.	157,736.	lung tumor hf-1785	0.429	0.754
851.	135,920.	Lung Tumor HF1602	0.346	0.608
851.	135,920.	Lung Tumor HF1651	0.592	1.041
851.	135,920.	Lung Tumor HF1729	0.611	1.074
851.	101,241.	Lung Tumor HF631	0.627	1.102
851.	101,241.	Lung tumor hf840	0.579	1.018
851.	157,736.	lung tumor R1057 vs epi pool	0.523	0.920
851.	157,736.	lung tumor R1094 vs epi pool	1.004	1.765
851.	157,736.	lung tumor R1094 vs epi pool	0.813	1.430
851.	157,736.	lung tumor R1372 vs epi pool	0.808	1.421
851.	157,736.	lung tumor R1372 vs epi pool	0.739	1.299
851.	157,736.	lung tumor R417 vs epi pool	0.431	0.758

851.	157,736.	lung tumor R542 vs epi pool	0.454	0.798
851.	157,736.	lung tumor R543 vs epi pool	0.651	1.145
851.	157,736.	lung tumor R544 vs epi pool	1.407	2.474
851.	157,736.	lung tumor R544 vs epi pool	0.655	1.152
851.	157,736.	lung tumor R685 vs epi pool	0.894	1.572
851.	157,736.	lung tumor R693 vs epi pool	0.836	1.470
851.	157,736.	lung tumor R693 vs epi pool	0.527	0.927
851.	157,736.	lung tumor R737 vs epi pool	0.479	0.842
851.	157,736.	lung tumor R742 vs epi pool	0.526	0.925
851.	157,736.	lung tumor R777 vs epi pool	1.041	1.830
851.	157,736.	lung tumor R777 vs epi pool	0.573	1.008
851.	157,736.	lung tumor R789 vs epi pool	0.849	1.493
851.	157,736.	lung tumor R789 vs epi pool	0.707	1.243
851.	157,736.	lung tumor R791 vs epi pool	0.869	1.528
851.	157,736.	lung tumor R791 vs epi pool	0.711	1.250
851.	135,920.	Lung Tumor RNA689	0.209	0.367
851.	135,920.	Lung Tumor RNA691	0.536	0.942
851.	135,920.	Lung Tumor RNA693	0.441	0.775
851.	157,736.	lung tumor vs epi pool	16.613	29.211
851.	157,736.	lung tumor vs epi pool	0.443	0.779
851.	157,736.	lung tumor	15.387	27.055
851.	157,736.	lung tumor/Ref.RNA	0.761	1.338
851.	157,736.	lung tumor1582/Ref.RNA	0.92	1.618
851.	157,736.	lung tumor1683/epipool	1.35	2.374
851.	157,736.	lung tumor1685/epipool	0.707	1.243
851.	157,736.	lung tumor1853/REF.RNA	1.491	2.622
851.	101,241.	lung tumorhf641	0.911	1.602
851.	157,736.	LungBaCa vs Epi	0.342	0.601
851.	157,736.	LungBAca1662 vs Epi	3.362	5.911
851.	157,736.	LungBAca	2.466	4.336
851.	157,736.	LungCaSq-hf1602	0.713	1.254
851.	157,736.	LungCaSq-hf1647	0.885	1.556
851.	157,736.	LungSqCa-hf1293	1.291	2.270
851.	157,736.	LungSqCa-hf1646	0.411	0.723
851.	157,736.	LungSqCa-hf?	1.112	1.955

Normal Lung Samples

851.	157,736.	normal lung 795	0.221	0.389
851.	157,736.	normal lung hf-1773	0.516	0.907
851.	157,736.	N. lung R1415 vs univ. ref	0.991	1.742
851.	157,736.	N. lung R1431 vs univ.ref	0.709	1.247
851.	157,736.	N. lung R417 vs univ.ref	0.669	1.176
851.	157,736.	lung Normal 1052 vs epi pool	0.603	1.060
851.	157,736.	lung Normal 1417 vs epi pool	0.317	0.557
851.	157,736.	lung Normal R417 vs epi pool	0.982	1.727
851.	157,736.	lung Normal R419 vs epi pool	0.335	0.589
851.	157,736.	423LungInflamTA1	0.517	0.909
851.	157,736.	488-551LungInflamTA1	0.486	0.855
851.	157,736.	488-552InflmdLungTA1	0.425	0.747
851.	157,736.	1054LungInflamTA1	0.56	0.985

851.	157,736. 1415LungInflamTA1	0.706	1.241
851.	157,736. 1415LungInflamTA1	0.494	0.869

GUIDELINES

Expression profiling — best practices for data generation and interpretation in clinical trials

*The Tumor Analysis Best Practices Working Group**

Microarrays are routinely used to assess mRNA transcript levels on a genome-wide scale. As use and acceptance increases, there is intensified focus on appropriate methods of data generation and interpretation, with important questions being asked about the best data analysis methods. The development of such 'best practices' is needed, as microarrays — in particular, Affymetrix oligonucleotide arrays — are becoming increasingly important in human clinical trials, both for differential diagnosis and monitoring of pharmacological efficacy. Here, representatives from high-volume microarray core centres consider the current status of 'best practices', focusing on the broadly used Affymetrix oligonucleotide arrays.

Microarrays represent a major technological advance in molecular biology. The introduction of any such advance is typically followed by a period of optimization and standardization. The latter is a crucial part of any maturing technology, as it allows an approach in which advances are made in parallel by individual researchers and companies who contribute new knowledge based on the existing standard. Any such standards must be constantly reassessed; stale or stagnant standards can inhibit the development of the technology.

Microarray-based mRNA-expression profiling can be considered to be the first mature genome-wide analysis technology, reflected in an increased interest in using microarrays as an endpoint in clinical trials. However, regulations of clinical trials require the development of clear standards for use and interpretation of microarray data (commonly referred to as quality control and standard operating procedures (QC/SOPs) and/or 'best practices'). Guidelines for reporting and annotation of microarray data from the Microarray Gene Expression Data (MGED) Society (see online links box) — using MIAME (Minimum Information About A Microarray Experiment) standards (BOX 1) and the MAGE-ML mark-up language[1,2] — represent an important step towards this goal. The efforts of this multinational academic–industry partnership has made it possible to develop databases that can house the many types of microarray data (see below) within the same data structure, enabling some data queries between experiments and experimental platforms. The ArrayExpress microarray database[3] (see online links box) is the first major publicly accessible database that adheres to this universal data-presentation platform, and some prominent journals (such as *Nature*, *Cell*, *EMBO Journal* and *The Lancet*) now demand that published microarray data conform to the MIAME standards. In addition, microarray manufacturers, such as Affymetrix, have implemented MIAME-compliant data output in their new software releases.

The MGED Society has effectively developed data-reporting guidelines, but it has not addressed issues of data generation and interpretation. The latter are more intimately coupled to the specific experimental platform. Of the three commonly used types of microarrays (spotted cDNA, spotted oligonucleotide and Affymetrix arrays), each has distinct methodologies associated with them; accordingly, the issues of data interpretation are also different (BOX 2). These differences make it difficult or impossible to develop cross-platform guidelines for data generation and interpretation. Best practices for spotted cDNA arrays are especially problematic because the manufacture of the arrays varies considerably from place to place. In addition, all spotted arrays use co-hybridization of a test RNA sample labelled with one colour FLUOROPHORE with a control RNA labelled with a different colour to which the test is compared on the same spot. The output is in the form of a ratio of hybridization signals that is comparable to other experiments only if the same control RNA is always used. Therefore, the development of standards in spotted arrays would require all laboratories to use the same control RNA solution before data could be easily compared.

Manufactured oligonucleotide arrays (both mechanically spotted and synthesized *in situ*) have the advantages of being centrally produced under controlled conditions. Affymetrix PHOTOLITHOGRAPHY-produced arrays have been available for nearly 10 years, whereas mechanically spotted oligonucleotide arrays have only very recently begun to appear in the marketplace. For example, Agilent Technologies (see online links box) recently released 17,000 60-mer oligonucleotides printed five times each on glass slides (85,000 FEATURES). Spotted oligonucleotide arrays typically have a single spot per gene (single probe measurement), whereas Affymetrix arrays provide multiple measurements — a series of independent or semi-independent oligonucleotides query each RNA in solution (the probe set) (BOX 2). Affymetrix probe sets are constructed from a series of perfect-match and paired-mismatch oligonucleotides, allowing some assessment of non-specific binding and performance of the probes. Overall, the Affymetrix probe sets provide a variety of measurements that allow robust measures of gene expression. The use of multiple perfect-match and mismatch probes for each gene enables the development of different methods of interpreting the hybridization patterns across the probe set and calculating a single 'expression level' or 'signal' that reflect the gene's relative expression level. A number of probe-set interpretation algorithms for Affymetrix arrays are available (see below for discussion).

> "[distinct methodologies] make it difficult or impossible to develop cross-platform guidelines for data generation and interpretation."

The increasing use of Affymetrix microarrays, and the emergence of this technology as an endpoint in clinical trials, has led to requests to develop, in both the pharmaceutical and academic research communities, best practices in data generation and analysis. Given the many differences between spotted cDNA, spotted oligonucleotide and Affymetrix arrays, the best practices need to be developed separately for each experimental platform; this is in contrast to data reporting that can be standardized across all platforms (BOXES 1 and 2). The Tumor Analysis Best

Box 1 | The MIAME guidelines for data reporting

The Microarray Gene Expression Data Society (MGED) is an international discussion group of microarray experts, with the primary goal of developing methods for data sharing between experimental platforms. The main output of this group has been the Minimum Information About A Microarray Experiment (MIAME) guidelines for microarray data annotation and reporting. The guidelines have been adopted by a number of scientific journals and have recently been endorsed for use by the US Food and Drug Administration and the US Department of Agriculture for pharmacogenomics projects.

The MIAME guidelines include descriptions of experimental design (number of replicates, nature of biological variables), samples used, extract preparation and labelling, hybridization procedures and parameters, and measurement data and specifications. These guidelines have been most important for the spotted cDNA and oligonucleotide experimental platforms (see BOX 2) in which the flexibility in microarray design and utilization also leads to considerable variation in array data generation and reporting between different laboratories. The guidelines do not attempt to dictate how experiments should be done, but rather provide adequate information associated with any published or publicly available experiment so that the experiment can be reproduced.

Box 2 | Microarray experimental platforms

There are three different types of microarray in common use: spotted cDNAs, spotted oligonucleotides and Affymetrix arrays.

Spotted cDNA arrays

Spotted cDNA arrays typically use sets of plasmids of specific cDNAs in gridded liquid aliquots. The inserts of each clone are typically amplified by PCR, and a few picolitres are physically spotted onto glass slides by liquid-handling robots. Robotic spotters can spot 100,000 spots per slide, and duplicate sets of clones are often spotted. The advantages of spotted cDNA arrays are that the content of each microarray is determined by the researcher, with complete flexibility in number and type of cDNA clones spotted. Also, the cost per array is relatively low, as the clone sets are a PCR-renewable resource and the glass slides are themselves inexpensive. The amount of the RNA that corresponds to each spot is determined relative to a second control RNA solution that is hybridized to the same spot, and a ratio is obtained.

Disadvantages of spotted cDNA arrays include the variable amount of DNA spotted in each spot, the 10–20% 'drop out' rate of failed PCR reactions or failed spots and mis-identification of clones (that is, the spot is not what you think it is). Also, there is no control over the actual sequence of the clone. As many gene-coding sequences contain regions of sequence that are shared with other genes, there are questions of specificity of the hybridization to the relatively large cDNA inserts. Spotted cDNA arrays were embraced by most academic centres, owing to their flexibility and relatively low cost.

Spotted oligonucleotide arrays

These arrays are also built by liquid handling on glass slides; however, the input solution is a synthetic oligonucleotide (often 60–70-mers). The resulting spotted material is typically of known concentration, of known sequence and is single stranded (all advantages relative to spotted cDNAs). Most of the process can be automated, leading to less sample mix-up and less drop-out of samples.

Disadvantages of spotted oligonucleotides include the relatively high cost of synthesizing large numbers of large oligonucleotides and the non-renewable nature of the resource. Spotted oligonucleotide arrays are becoming increasingly available.

Affymetrix GeneChips

These microarrays are factory designed and synthesized. Design is done using software to choose a series of 11 25-mer probes from the 3' end of each transcript or predicted transcript in the genome; each of the 11 probes is then paired with a similar mismatch probe that is designed to contain a mutation in the centre. The latter serves as a form of control for hybridization specificity. Synthesis of arrays is done using light-activated chemistry and photolithography methods, and feature size can be reduced to approximately 8 μm^2, with about 1 million probes in a 1.2 cm^2 glass area. Probe-set algorithms interpret the signals from each 22-oligonucleotide probe set, and derive a single value (signal) from the patterns of hybridization to the 22 individual probes. This signal is then normalized to the entire microarray, or to the probe sets across an entire project.

For a more general discussion of normalization and analysis methods of different microarray platforms, the reader is referred to the excellent web information resource of the MGED group (see The MGED Data Transformation and Normalization Working Group in online links box).

Practices Working Group (see BOX 3) was convened to discuss and develop best practices for Affymetrix microarrays, including QC and SOPs for both data generation and data analyses. The first meeting was held in Santa Clara in March 2003, followed by a series of conference calls that focused on discussions of data generation and analysis standards for the Affymetrix oligonucleotide arrays. The Working Group deliberately focused on a platform that has widespread usage and is most likely to be used in clinical trials owing to the previously standardized manufacturing process. Here, we discuss recommendations for experimental design, probe-set analysis algorithms, signal/noise assessments and biostatistical methods.

Experimental design

Appropriate experimental design is a key aspect of all science, and microarray studies are no exception. The relatively high cost of some commercial microarray platforms is a frequently cited reason for suboptimal experimental design, especially with regards to the number of replicates. Data interpretation is inevitably compromised when replicates are decreased.

Replication in cross-sectional studies. The appropriate number of microarray replicates for any particular condition or time point depends on the source of biological variability in the study samples. Inter-individual variability is very large in outbred (genetically heterogeneous) humans, but is very small within inbred mouse strains. For example, expression profiles derived from muscle from different mice are not more variable than from muscles isolated from one mouse[4]. Defining the confounding variables that contribute to experimental variability, such as intra-subject, inter-subject, inter-group and technical variation (microarray protocol), is needed to design and statistically power a study, and to determine the number of replicates that are needed. In general, inbred mice require testing only three or four mice per group. We and others have found that five or six out-bred rats per group provide statistically robust results[5,6]. By contrast, human samples require considerably more individuals per group. Key variables in human samples include tissue heterogeneity, stage of disease and inter-individual variation, all of which have been found to be major confounding variables[7].

Replication in longitudinal studies. It has long been recognized that, in human clinical trials, LONGITUDINAL DESIGNS provide considerably greater power at lower numbers of replicates.

They best control for inter-individual variation because each subject serves as their own control. Serial blood sampling from single subjects is the least invasive[8] (see below for further discussion), and, for example, cancer patients are often longitudinally sampled[9]. Serial biopsies of other tissues are more invasive: however, a number of serial human muscle biopsy studies of healthy volunteers after different types of exercise training have begun to appear in the literature[10,11].

Expression profiling of blood samples (longitudinal or CROSS-SECTIONAL DESIGN) is the protocol that is most likely to be used in human clinical trials. One of the Working Group's goals was to establish SOPs for blood sample collection and RNA isolation in clinical trials. A specific follow-up report of these recommendations will be published elsewhere. Such a protocol must be easily adaptable to multiple trial sites, with relatively little need for resident expertise to carry out the isolation protocol. So far, standard methods for isolating peripheral-blood mononuclear cells have shown the most reproducibility, although others are being tested (see Affymetrix Technical Note in online links box). Cells isolated soon after collection can be flash frozen for storage and subsequent RNA isolation or an RNA stabilizing compound can be added if the samples need to be transported.

Tissue/cell heterogeneity. Tissue heterogeneity is a major confounding variable in most microarray experiments. In inbred mice, tissue heterogeneity is typically normalized by using whole organs. This is rarely possible in human experiments, and particularly not in clinical trials; the limited amount of human tissue that is available exacerbates heterogeneity. The mixed cell populations of peripheral blood can be thought of as a tissue heterogeneity problem similar to that encountered in all solid tissue and tumour biopsies. Indeed, a recent study showed that variation as a result of tissue variability in human muscle biopsies often exceeded inter-individual variability[12]. One potential solution to the tissue heterogeneity problem lies in bioinformatic methods. If computer software can be trained to recognize the expression profiles of each individual cell type within a mixed tissue sample, then it should be possible to subtract them from each other and to renormalize to obtain a set of cell-specific expression profiles derived from a single mixed profile. This will be most easily done on tumour biopsies, in which the main cells of interest are tumour versus contaminating normal tissue. Although there are no published examples so far, such methods are maturing rapidly.

Box 3 | **The Tumor Analysis Best Practices Working Group***

The Tumor Analysis Best Practices Working Group is a group of investigators who study the best practices of tumour analysis in humans taking part in clinical trials. The following authors are members of the Group:

- Eric P. Hoffman is at the Research Center for Genetic Medicine, Children's National Medical Center, Washington DC 20010, USA. email: ehoffman@cnmcresearch.org
- Tarif Awad, John Palma, Teresa Webster, Earl Hubbell and Janet A. Warrington are at Affymetrix, Santa Clara, California 95051, USA. emails: tarif_awad@affymetrix.com; john_palma@affymetrix.com; teresa_webster@affymetrix.com; earl_hubbell@affymetrix.com; janet_warrington@affymetrix.com
- Avrum Spira is at The Pulmonary Center, Boston University Medical Center and the Bioinformatics Program, Boston University, Boston, Massachusetts 02118, USA. e-mail: aspira@lung.bumc.bu.edu
- George Wright is at the Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institute of Health, Bethesda, Maryland 20892, USA. e-mail: wrightge@mail.nih.gov
- Jonathan Buckley and Tim Triche are at the Children's Hospital, University of California, Los Angeles, California 90089, USA. e-mail: buckley@hsc.usc.edu; triche@hsc.usc.edu
- Ron Davis, Robert Tibshirani and Wenzhong Xiao are at Stanford University, Palo Alto, California 94303, USA. e-mails: dbowe@stanford.edu; tibs@stat.stanford.edu; wzxiao@pmgm2.stanford.edu
- Wendell Jones is at Expression Analysis Inc., Durham, North Carolina 27713, USA. e-mail: wjones@expressionanalysis.com
- Ron Tompkins is at Harvard University, Boston, Massachusetts 02115, USA. e-mail: rtompkins@partners.org
- Mike West is at the Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27708, USA. e-mail: mw@stat.duke.edu

An experimental alternative to mitigate confounding tissue heterogeneity is to isolate pure cell populations for expression profiling. Many such methods are well developed in the research laboratory, including FLUORESCENCE-ACTIVATED CELL SORTING (FACS)[13], NEGATIVE CELL ISOLATIONS from blood (for example; Stem Cell Technologies RosetteSep)[14] and LASER CAPTURE MICRODISSECTION[15]. To research scientists, the profiles that are derived from isolated cell types are a more intuitive approach to define biologically relevant pathways. However, it should be noted that uses of array-based analysis of gene expression approved by the US Food and Drug Administration (FDA) will probably focus on reproducibility and robustness (as well as on predictive accuracy), rather than on biological interpretation or justification. The high-tech methods used to isolate specific cell types from clinical samples are unlikely to make their way into clinical trials unless tissues are procured in a highly centralized way.

"If computer software can be trained to recognize the expression profiles of each individual cell type within a mixed tissue sample, then it should be possible to subtract them from each other..."

Procedural variation. Beyond the usual issues of sampling and accrual, gene-expression data will be subject to many additional sources of error. For example, the surgical removal and processing of tumour tissue can vary considerably from site to site. Laboratory QC procedures in tissue handling, RNA extraction and processing, and variations in protocols for data management and processing will need to be addressed in any clinical trial design. In particular, prolonged tissue ISCHAEMIA prior to processing of surgically RESECTED tissue can significantly alter gene expression[16]. All tissue samples should be flash frozen within minutes of surgery and stored at –80°C or below. Samples should also be kept in small, airtight containers and kept from drying out during frozen storage by placing fragments of ice in with the sample.

Technical variability

The standard laboratory protocol for generating RNA profiles using Affymetrix microarrays involves a series of steps (FIG. 1).

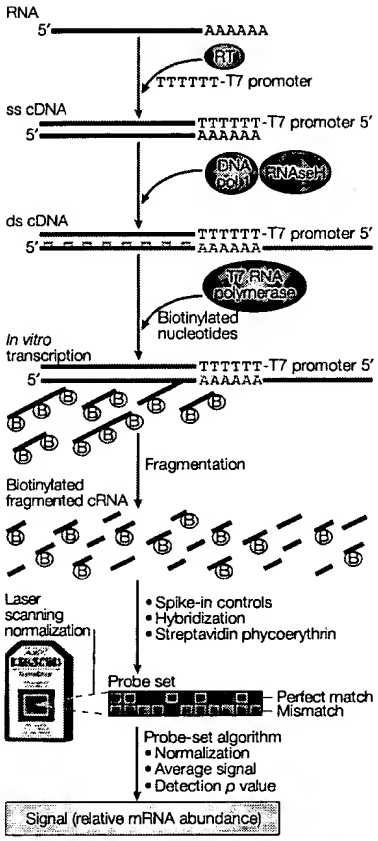

Figure 1 | **Sample processing and microarray interpretation of Affymetrix GeneChips.** Flash-frozen tissue (~50 mg) is homogenized to isolate total RNA. Single-stranded (ss) cDNA and then double-stranded (ds) cDNA is made from ~5 μg of total RNA. Double-stranded cDNA contains a T7 RNA-polymerase promoter adjacent to the 3′ polyA tail of each transcript. It is transcribed *in vitro* to generate more than 400 biotinylated cRNA molecules for each ds cDNA molecule. The biotinylated cRNA is fragmented and hybridized to the microarrays. Each transcript is queried by one or more probe sets of 11 perfect-match and 11 paired-mismatch oligonucleotides (the latter contain a centrally located point mutation as a form of hybridization specificity control). Currently available Affymetrix microarrays have ~54,000 probe sets on each 1.28 cm^2 glass microarray (~1.2 million 25-mer oligonucleotides on the HG-U133Plus 2.0 array). The biotinylated cRNA fragments hybridize to the appropriate oligonucleotide features. A laser scanner determines the amount of bound biotinylated cRNA indirectly through the streptavidin-conjugated phycoerythrin fluorescence at each feature within a probe set. The component probe pairs are interpreted and averaged to arrive at a single signal that reflects the relative abundance of the original mRNA. Probe sets are interpreted by any one of a number of probe-set algorithms, each providing a signal that reflects the relative hybridization intensity across the probe set. RT, reverse transcriptase.

RNA isolation. RNA quality and quantity is crucial to the success and reproducibility of the expression profiles. RNA quantity and quality is generally checked by complementary methods: UV 260/280 ratio >1.8, agarose gel electrophoresis or an Agilent Bioanalyzer to visualize clear 18S and 28S ribosomal RNA bands. Total RNA (5–10 μg) is input into the cDNA/cRNA reaction, with an expected corrected yield of biotinylated cRNA of between 4- and 10-fold greater than the total RNA input (so 5 μg of total RNA must yield at least 20 μg of biotinylated cRNA, or the sample is discarded). The biotinylated cRNA should be 500–3,000 base pairs (bp) in size. After fragmentation, the cRNA should be 50–200 bp. The Working Group recommends that samples that do not meet these criteria should be discarded.

If RNA amount is limiting — as is the case, for example, with laser capture microscopy samples, flow-sorted cell samples or small tissue samples — a two-round amplification protocol can be used. For example, 200 ng of total RNA is processed for *in vitro* transcription (IVT), with the same goal of 4–10-fold amplification (>800 ng of cRNA output). One hundred nanograms of this cRNA is then reverse transcribed into cDNA using random primers, after which a second IVT is done. The second round IVT should result in a 400-fold amplification.

Microarray controls. Hybridization controls include visualization of the image so that any abnormalities in hybridization patterns can be detected. ProbeProfiler from Corimbia Inc. is a program with extended capabilities for detecting defects in microarray manufacture. Affymetrix MAS 5.0 software adjusts the microarray-scanned image to a common target intensity by using a scaling factor. In addition, a general index of chip background and noise is represented by the percentage of 'present calls' (probe sets for which the hybridization to the perfect-match probes is significantly higher than mismatch hybridization). The Working Group believes that both the scaling factor and the percentage of present calls are important QC criteria. Considering MAS 5.0 chip analyses, the scaling factors to normalize chips within a project should lie within two standard deviations of the mean, with present calls being greater than 25% (BOX 4). The percentage of present calls is often lower when B or C arrays that contain higher proportions of more poorly characterized transcript units (expressed sequence tags or computer-predicted open reading frames) are used. The percentage of present calls across a set of samples should be consistent, within a range of 10%. Some software packages allow the identification of statistical 'outlier' microarrays in a group of microarrays in a given project, which additionally enables the experimenter to flag and exclude specific microarrays that are not acceptable for an analysis. In addition to these criteria, acceptable hybridizations must have adequately intact input RNA as shown by 3′ to 5′ ratios of hybridization within probe sets. A typical control is the glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) gene, which should have 3′ to 5′ ratios of less than 3 (BOX 4).

The QC criteria provided above are based on MAS 5.0 probe-set algorithms and data analyses. The measures of present calls and scaling factors are useful and serve as initial summary measures of the performance of a particular microarray. However, more focused statistical methods, coupled with routine visual inspection of images, hold promise for the continuing improvement of data quality and screening abilities.

Large-scale analyses of microarray data across laboratories have not yet been reported. However, the Working Group feels that adherence to the above QC criteria, using standard RNA isolation and processing methods, should yield data that are consistent between laboratories and intrinsically comparable. The same set of criteria can also be used as best practices for data generation in the design and conduct of clinical trials.

Standard clinical laboratory practice is to develop programmes for submission of known samples to different laboratories and assessment of comparability of results. Such programmes are under development within larger collaborative efforts, such as the National Heart, Lung and Blood Institute (NHLBI) Programs in Genomic Applications (see the HOPGENE Program for Genomic Applications in online links box) and the National Institute of General Medical Sciences (NIGMS) Glue Grant (see online links box).

Data analysis and interpretation

Signal generation versus statistical analyses. Two relatively distinct steps underlie all data analyses of Affymetrix oligonucleotide microarrays: the development of a normalized 'signal' for each transcript on each microarray and the subsequent statistical analysis of differences in signals between different arrays. The first step involves probe-set algorithms that use all, or part, of the component signals within a probe set and then derive a single signal that is representative of the relative abundance of each mRNA queried in each array. The second step is the

application of bioinformatic and statistical methods to identify interesting subsets of the assembled data of all arrays within a project. There is considerable debate about the best methods for both of these steps (see below for a discussion). Although the two steps are separable, it is clear that they have a marked influence on each other. It is in this realm that the bioinformatics of microarrays becomes avant-garde, and with the groundbreaking nature of research comes considerable debate as to what is appropriate in any specific situation.

Before discussing the different methods for probe-set analysis and data interpretation, it is important to point out that much of the debate in the field of bioinformatics about microarray interpretation revolves around signal/noise ratios. A common assumption is that signal/noise ratios across a microarray are homogeneous, or at least similar in magnitude. This might be true for general background hybridization, but not for the performance of probe sets. In any particular microarray, there are probe sets that give very strong and clear hybridization patterns and those that perform poorly. Many of the best performing probe sets (those with a highly significant probe-set detection *p* value) reflect highly expressed transcripts with no closely related sequences that might cross-hybridize. Low-level transcripts, or transcripts that belong to gene families with highly homologous sequences derived from distinct genes, often have corresponding probe sets that do not perform as well and might have a significant, if not overwhelming, noise component. The signal from such probe sets is difficult to interpret, and data interpretation can be limited to only the best performing probe sets, although arguably the most interesting data comes from the genes that are expressed at low levels but that still show significant differences between samples.

Determining adequate sensitivity of the signals and signal/noise responses relative to the absolute quantity of mRNA in clinical samples is crucial as microarrays become a component of clinical trials and diagnostic models. Affymetrix arrays provide a concentration of each mRNA queried relative to the genome-wide mRNA profile of the sample; it is assumed that the global mRNA content of a tissue as a whole does not change significantly, making relative mRNA quantification an accurate reflection of the response of the individual gene. This method differs from absolute quantification of specific mRNAs (such as S1 NUCLEASE PROTECTION and REAL-TIME PCR), or the isolated transcript ratio determined by co-hybridization of two samples to spotted cDNA or oligonucleotide arrays (BOX 2).

Box 4 | Quality Control metrics for Affymetrix microarrays

RNA quality
Optical density 260/280 of 1.8–2.1 | Agilent Bio-Analyzer | Gel electrophoresis

cDNA/cRNA efficiency
>4-fold amplification from total RNA | 500–3,000 bp prior to fragmentation | 50–200 bp after fragmentation

Chip hybridization
Image inspection for defects | Scaling factors within two standard deviations within a project | MAS 5.0 present calls >25% for the A-SERIES ARRAYS, including the HG-U133Plus 2.0 array | Percentage present calls for the B- AND C-SERIES ARRAYS are typically lower | 3′/5′ *GAPDH* ratios <3

Project normalization
The detection of statistical outliers for chips, probe sets or individual probe pairs requires normalization and analysis across an entire project. This is afforded by the dCHIP and ProbeProfiler, and other software packages. Data-analysis packages that rely on intra-chip normalization and scaling typically do not enable detection of statistical outliers.

Chip outliers
Probe-set outliers | Probe-pair outliers | Range in present calls <10%

Affymetrix arrays achieve considerable sensitivity through the inherent redundancy of the probe set; however, the Working Group acknowledged that some genes, such as some cytokines that are functional at very low expression levels, are probably below the limit of detection.

The Working Group agreed that each project will have its own signal/noise optimum, and analysis methods that prove best for one project might prove unsuitable for another. Ideally, a signal/noise ratio should be optimized for each project or trial using different probe-set algorithms and data-filtering methods, and some systematic efforts towards this end are beginning to appear in the literature[17].

"... adherence to the above QC criteria ... should yield data that are consistent between laboratories and intrinsically comparable."

After a signal is derived for each probe set, data is interpreted using statistical and visualization methods. All statistical methods run into two generic problems when faced with microarray data that are inter-related. The first is the curse of dimensionality — each gene is potentially related to every other gene, so all permutations of all available data must be considered, leading to an exponentially increasing number of possible associations in multidimensional space. The problem arises when associations (samples) become lost as the dimensionality increases — associations lose their local value and become generically global in statistical terms. Statistical models attempt to circumvent this curse by requesting larger and larger sample sizes, but fulfilling the requests becomes functionally impossible for the experimentalist. There is no easy answer to these problems and they remain a challenge for future bioinformatics research that uses microarrays[18].

Derivation of signal: probe-set algorithms and normalizations. One of the key advantages of the Affymetrix platform is the multiple measurements that are intrinsic to the probe set — most probes include 11 perfect-match and 11 paired-mismatch 25-bp oligonucleotides per gene (FIG. 1). Previous versions of GeneChip arrays used probe-set design methods that led to considerable overlap between probes, so that hybridizations to each feature/probe were not independent measurements; this led to considerable uncontrolled weighting of the contribution of any particular region of sequence to the resulting signal. All recent chips use a much more refined probe-set design with less overlap and considerably better performance of the probe set. Improvements in array and probe-set designs have been accompanied by an evolution in primary analysis algorithms and the supporting software provided by Affymetrix for data analysis and interpretation[19]. Affymetrix default algorithms are based on well-documented statistical methods, namely the robust TUKEY'S BI-WEIGHT ESTIMATOR and WILCOXON'S SIGNED RANK, to calculate the final probe-set signals and associated *p* values, respectively[19,20].

Table 1 | **Comparisons of probe-set analysis algorithms**

Algorithm	Penalty for mismatch signal	Normalization method	Outlier detection and correction	Sensitivity*	Specificity‡
Affymetrix MAS 5.0	High	Individual chips	Little	Good	Excellent
dCHIP difference model	High	Cross-project	Moderate	Good	Excellent
dCHIP	None	Cross-project	Moderate	Excellent	Good
RMA	None	Cross-project	Moderate	Excellent	Good
ProbeProfiler	Moderate	Extensive	Extensive	Good	Good

*Sensitivity is based primarily on ROC (receiver operating characteristic) curves of spike-in mRNA data based on published reports (see http://www.bioconductor.org)[21,23].
‡Specificity measurements are based both on expectations from mismatch weights and published observations in experimental data sets[17,18].

Affymetrix has announced plans to continue to improve the software components of the GeneChip platform. The upcoming release of the GeneChip Operating System (GCOS) is expected to incorporate refinements in the user interface, data management and analysis algorithms. Software tools aside, the most significant development on the analysis front is arguably the decision by Affymetrix to release previously proprietary chip-design details, such as probe sequences, chip-design parameters and file APIs (applications programming interfaces). The goal is to encourage scientists to develop innovative analysis tools that can potentially derive more biological value from GeneChip expression data. The challenge of providing a constantly growing and evolving body of information associated with arrays has been solved in part with a web-based tool. The company's NETAFFX web site (see online links box) serves as the public portal for detailed information on chip design and has become a valuable resource for biological follow-up of GeneChip expression results. Third-party software developers can find additional support, including information on file APIs, through the Affymetrix Developers' Network (see online links box).

Encouraged in part by the openness of the platform and spawned by an increase in knowledge and experience in array data analysis, scientists are developing a number of alternative algorithms for probe-set analysis, with the goal to derive the best signal that is representative of the mRNA level for each gene. As each signal is relative to other signals in the experiment (both between arrays for the same gene and relative to all other genes on the array), the process of normalization is intimately tied to derivation of signal. The more commonly used alternative probe-set analysis algorithms include dCHIP[20], RMA[21] and ProbeProfiler (TABLE 1).

It is outside the scope of this article to discuss the nature of the different probe-set interpretation and normalization algorithms in depth, and the reader is referred elsewhere[22]. The algorithms differ in a number of important ways (TABLE 1). First, the PENALTY WEIGHT that is assigned to the mismatch probe varies — MAS 5.0 assigns a relatively heavy penalty for cross-hybridization to the mismatch probe, RMA assigns no weight and dCHIP gives the choice of providing weight or no weight. Second, the ability to discard outlier signals varies from package to package, with dCHIP and ProbeProfiler having refined methods to detect outliers at each level of analysis (probe, probe set and microarray). These packages are able to replace deviant probes with expected data based on the remainder of the probe set, and/or flag abnormal probe sets and arrays for possible exclusion from further data analysis. Third, the method of normalization varies from within a single array (MAS 5.0) to a project-based normalization (dCHIP, RMA and ProbeProfiler). Finally, MAS 5.0 provides a detection *p* value, in which a number is assigned to the confidence of the signal in question. This can be used to weight different probe-set signals in subsequent data interpretations.

> "...robust feature selection for the purpose of diagnosis and molecular markers in clinical trials requires robust statistical methods..."

The output of all packages is a normalized signal (with or without an associated detection *p* value) for each probe set on each array. These signals are then fed into data interpretation packages for statistical analyses and data visualizations.

Different probe-set interpretation algorithms lead to different results. Members of the Working Group often encounter ~50% concordance in general data output in their own work between comparisons of two different algorithms. However, it is crucial to note that the large majority of discordant data lies in regions of relatively poor signal/noise ratios, and concordance deteriorates in experiments with high levels of confounding noise. In general, the programmes that put less weight on the mismatch show better sensitivity (linearity) when signals are noisy (TABLE 1). However, this increased sensitivity can come at a cost of substantial contaminating noise[17].

The Working Group recommends using at least two probe-set algorithms for comparison and prioritization of gene selection (for example, MAS 5.0 and the dCHIP difference model).

Data interpretation. Most published microarray papers could be considered data-poor in terms of replicates and systematic statistical analyses, but data-rich both in terms of amount of high-quality data generated and significant research findings. Below, we point out the most appropriate current bioinformatics methods and additional methods that require further development so that data can be more fully mined for information content.

A second general backdrop to the following discussion is that data visualization is one of the most powerful data interpretation tools, yet it rarely obeys statistical principles. The resolution of the human eye, coupled with the abstract computational power of the human brain, lies behind the popularity of hierarchical clustering and other non-statistical principles and visualization methods. However, the eye and brain are poorly suited to spontaneously deriving statistical support.

There are two general types of experimental design that lend themselves to different types of statistical and visual analysis: the cross-sectional study and the TIME-SERIES STUDY. The cross-sectional study typically has gene or pattern selection as the goal: the identification of one or more genes or patterns of expression that are diagnostic of the condition or state

under study. This 'gene selection' might be for truly diagnostic purposes (for example, differential diagnosis of leukaemia), or might be intended to identify relevant biochemical pathways. In both cases, the gene or pattern selection must be robust, usually implying a statistically principled approach, with subsequent validation by predictive computer modelling (internal cross-validation) or, preferably, prospective validation on new data.

Feature selection can be the main limiting factor in evaluation of the predictive performance of an analysis method when there are many predictors to select from. This was a 'mantra' for some of the senior statisticians involved in predictive modelling with gene-expression array data for several years, but only now do the non-statistical users and developers of predictive models from non-statistical perspectives begin to appreciate these issues. Proper validation of any model or algorithm that relies on explicit feature selection — such as choosing a subset of 70 genes from 20,000 — that underlies the resulting prediction simply must ensure that the analysis is tested by internal cross-validation that includes feature re-selection as part of the validation[23,24]. The Working Group acknowledged that prospective validation of any findings using new data is the acid test of predictive performance. The focus on feature or gene selection is vitally important when microarrays are used for differential diagnosis and has been best studied in cancer biopsy/tissue studies.

An increasing proportion of microarray studies focus on delineation of biochemical pathways that are modulated in response to some stimulus. In practice, these studies typically use feature selection to identify potential pathways that are involved in the response of the cells or tissues. Validation is then done on the identified biochemical pathways of interest, using mRNA (real-time PCR) or protein studies, often proving cause and effect in experimental models.

The Working Group notes that robust feature selection for the purpose of diagnosis and molecular markers in clinical trials requires robust statistical methods, as outlined below, and the burden of proof lies with statistical validation. For microarray experiments designed to delineate biochemical pathways, feature selection is used for generating a hypothesis and the burden of proof of the hypothesis lies with laboratory-based research, often at the protein level.

For feature selection, the Working Group recommends that users experiment with various statistical methods (such as standard parametric tests, nonparametric methods, false discovery rate and related methods[25], global or local shrinkage of raw signal intensities and Stanford's 'nearest shrunken centroids'[26]). Developments related to SURVIVAL DATA ANALYSIS are receiving increased attention because clinical trials will raise the need to move that way. As a corollary, analysis methods that focus on signatures of groups of genes (such as averages of clusters, Duke's metagenes[27–29] and Stanford's eigengenes[30]) seem worth stressing in predictive contexts.

Glossary

A-, B- AND C-SERIES ARRAYS
A series of human, rat and mouse Affymetrix arrays released in 2003, in which the A array contained the best-characterized genes, and B and C arrays contained less well-defined expressed sequence tags. In 2004, all probe sets have been condensed so that there is only one microarray per species that covers the entire genome.

CROSS-SECTIONAL DESIGN
The use of different subjects in an experimental and control group or groups. The statistical analysis compares the median and variation within each group relative to the other groups.

FEATURE
Typically one element (spot) on a microarray. In spotted cDNA or oligonucleotide arrays, features correspond to genes or transcripts; in Affymetrix arrays, there are typically 22 elements per probe set and often multiple probe sets per gene, so a feature might refer to a single oligonucleotide, a probe pair or a probe set, or a gene with multiple probe sets. In bioinformatics it is most often synonymous with a gene.

FLUORESCENCE-ACTIVATED CELL SORTING
(FACS). A method whereby dissociated and individual living cells are sorted, in a liquid stream, according to the intensity of fluorescence that they emit as they pass through a laser beam.

FLUOROPHORE
A small molecule, or a part of a larger molecule, that can be excited by light to emit fluorescence.

ISCHAEMIA
The loss of blood supply, and hence oxygenation, to a tissue or cells.

LASER CAPTURE MICRODISSECTION
A technique in which individual cells, or regions of tissue, are excised from a histological preparation, using specially equipped microscopes, and isolated for further study.

LONGITUDINAL DESIGN
The use of multiple samples from the same subject. With this design, each subject serves as their own control, eliminating confounding inter-individual variations at baseline; paired *t*-tests are used to interpret the data.

NEGATIVE CELL ISOLATION
The use of antibodies or other reagents to remove all unwanted cells from a mixed population of cells. In this method, the desired cells are not exposed to bound antibodies, thereby avoiding potential activation or other molecular alteration in the desired cells.

PENALTY WEIGHT
In Affymetrix arrays, hybridization to the 'mismatch' probe of a probe pair might or might not be considered as a form of measurement of noise or background, and can be factored into the signal seen with the paired 'perfect match' as a penalty weight.

PHOTOLITHOGRAPHY
The process of using light to either etch or activate regions of a surface (substrate). This method is used in microelectronics to create integrated circuits and processors.

REAL-TIME PCR
The quantification of the amount of PCR product during each cycle of a PCR reaction. The product concentration, as a function of cycle number, provides a good estimation of the relative quantity of the mRNA being tested.

RESECTION
Surgical removal of tissue, most commonly used for removing tumorous masses from surrounding tissue.

S1 NUCLEASE PROTECTION
An experimental method for determining mRNA transcript concentration in a tissue or cell RNA sample. It involves using labelled DNA probes that bind the RNA, with overhanging non-hybridized tails of the probe then being digested by the S1 nuclease. This creates a smaller labelled DNA probe that is indicative of the abundance of the mRNA being tested.

SURVIVAL DATA ANALYSIS
A battery of statistical methods applied to data when mortality is often the only, or best, measured outcome.

TIME-SERIES STUDY
The use of a series of samples taken at defined time points after a defined stimulus. In mice and rats, the samples at different time points are usually from different animals. In humans, time-series studies are necessarily longitudinal to avoid additional confounding noise.

TUKEY'S BI-WEIGHT ESTIMATOR
Many statistical tests require underlying definitions that are assumed to be valid (for example, tumour versus non-tumour), and require data that show a normal distribution. Microarray data, and the clinical information underlying the definition of samples, is often less exact, with genes or samples often performing as statistical outliers. Tukey's bi-weight estimator is one of the M-class of statistical models that is less sensitive to outliers and performs more gracefully when underlying assumptions are inexact.

WILCOXON'S SIGNED RANK
A statistical test that investigates the population median of paired differences. It is well suited for microarray work as it treats each gene as an independent variable and does not require normal distributions of the data.

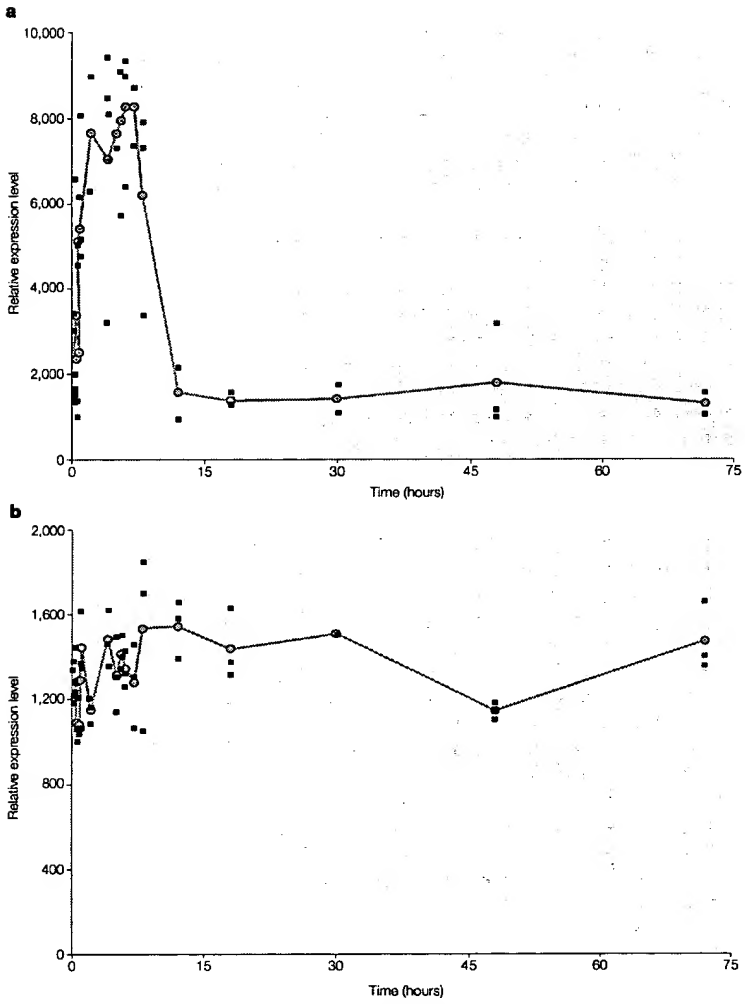

Figure 2 | **Dense time-series data with adequate replicates can provide robust visual interpretation of data.** Dynamic, single-gene queries can provide visually compelling results and avoid many issues that complicate statistical analyses of cross-sectional microarray studies. Dynamic queries of the *got1* transcript probe set from a time-series study is shown. The *x*-axis represents time (in hours) after rats were given a bolus of 3-methylprednisolone. The *y*-axis represents relative expression level. Individual rats were studied at each time point (green data points), with both liver (panel **a**) and muscle (panel **b**) tissue taken from the same animals. Averages of the replicates for each data point are shown (in magenta); the graph line is drawn between the averages. At baseline (time 0), the *got1* transcript has a normalized signal of about 2,400 in liver and 1,200 in muscle. The gene is clearly responsive to 3-methylprednisolone in liver (panel **a**), where expression rapidly increases within the first hour, plateaus between 1–7 hr, then quickly falls back to baseline by 12 hr. The replicates appear relatively consistent. By contrast, the same gene in muscle does not seem to respond to the drug; the variability in replicates is larger than any temporally-relevant change. Data from http://microarray.cnmcresearch.org/singlegenemain.asp and REFS 6,32.

Whatever the specific statistical model that is applied for prediction, using aggregate gene expression has important consequences: measures of aggregation of expression over a group of genes with related profiles can reduce dimension (thereby mitigating the curse of dimensionality). This can reduce multiplicities and, to some degree, ease the problems of gene selection, multiple testing and co-linearity, while improving signal estimation by averaging correlated noise components.

Data visualizations, time series and candidate genes. The above discussions of biostatistics all assume that the analysis is targeted towards a cross-sectional study, in which the primary goal is diagnostic gene discovery (gene or feature selection). In other words, a series of microarrays with a very large number of transcripts defines the very small minority of genes that are correlated and therefore predictive of the biological variable of interest. There are alternatives to this standard experimental design that use entirely different types of analysis, and the statistical issues are also quite different, as explained below.

The time-series study, if done with enough time points, can provide an effective antidote to the curse of dimensionality — the action of any gene during a time-series study should make biological sense, such that each signal is relatively easily discernible from noise. Visual query of a large time-series data set for single gene responses to the controlled variable either might meet expectations and is therefore valid, or might not meet expectations and is discarded as uninteresting. As an example, we show a time-series study in which rats are given a bolus of methylprednisolone, after which their liver and muscle are studied as a function of time (FIG. 2). In this case, the same gene (*got1*) is queried using a web-based dynamic visualization tool, first in liver (FIG. 2a) and then in muscle (FIG. 2b). The data in the top panel are visually compelling; *got1* in liver responds quickly and strongly to a bolus of 3-methylprednisolone, with relatively consistent replicates (each data point comes from a different animal) and a time course that is visually assuring so that complex statistical tests of the transcriptional response as a function of time are not needed. On the other hand, the same gene in muscle does not seem to respond to the drug[6,31] (FIG. 2b). Through such gene queries, the variability in replicates and the appropriateness of the action of the gene as a function of time can quickly be assessed. Another advantage of time-series data is that such profiles act as biomarkers that are amenable to analysis and interpretation using pharmacodynamic models that predict the underlying mechanisms of control of gene expression[32].

The Working Group agreed that data-rich, time-series experimental designs provide some latitude in reporting significant findings and that the query of individual genes within large data sets can circumvent complex issues of multidimensionality of data.

Future areas of development

The data-rich and highly dimensional nature of microarray data serves as a model for future dissection and understanding of biological

systems in general, including proteomics and integration of mRNA profiling and proteomics. The Working Group discussed data analysis needs within the microarray community and agreed that, along with the incorporation of QC, SOPs and optimized or customized signal/noise analyses in initial project signal generation, the back-end statistics needs to reach a commonly accepted method of dealing with the curse of dimensionality before microarrays can be reliably used in clinical trials. Statisticians need to focus more on representation of prediction results in terms of probabilities and associated measures of uncertainty, and reach a consensus on what is acceptable. In the meantime, it is likely that specific marker or diagnostic genes will be extracted from pilot profiling studies, and then only this small subset of genes will be used as a clinical trial endpoint. This data limitation approach removes much of the curse of dimensionality, but is liable to ignore the large majority of data, thereby decreasing the potential power of the study and bringing into question the use of microarrays in clinical trials.

A move towards the standardization of reporting of prediction accuracy would be desirable when assessing predictive accuracy through within-sample cross-validation. The Working Group suggests that one or more validation techniques be used when reporting predictive genes: leave-one-out and 10% cross-validation summaries, or true validation data sets. Communicating uncertainty about predictive performance is also key and will help evaluate results based on varying sample sizes. The Working Group suggests that until this information is routinely presented in published papers, it will be difficult to reach an acceptable consensus for use in clinical trials.

Conclusions

There are four key areas of optimization and standardization that are largely independent: study design, technical variability (QC/SOP of data generation), analysis method variation (signal/noise optimization using probeset algorithms and normalizations) and back-end statistical analyses. Statistics of clinical trial design is crucial: gene-expression data does not mitigate the need for sound and relevant design and analysis, nor does it challenge what we know about design. The field is quickly maturing from the small-chip-number hit-and-run type projects to those with a more robust study design. However, study design depends ultimately on appropriate powering of a study, which is greatly affected by both the chip-analysis algorithms that are used and the biostatistical data analysis.

Development of back-end statistical methods for data representation/summary and for high-level analysis remains an active area of research for both academic and commercial users, and is likely to remain so in the near future. We are some way from defining standards of summary signal intensities alone and even further from considerations of standardization of analytical methods for inference and prediction in clinical contexts. In regulated clinical studies, such standards will be enforced partly by the US FDA as submissions of medical test/device protocols emerge and increase in number. Even then, however, many approaches to data analysis and modelling will be used and developed, which is, of course, to be supported. It is very difficult to influence the research community, especially when the variety of problems that are encountered promotes the need for refined and new approaches.

*Members of The Tumor Analysis Best Practices Working Group are listed in Box 3.
Correspondence to Eric P. Hoffman at the Center for Genetic Medicine, Children's National Medical Center, 111 Michigan Avenue NW, Washington DC 20010, USA.
e-mail: ehoffman@cnmcresearch.org

doi:10.1038/nrg1297

1. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. *Nature Genet.* **29**, 365–371 (2001).
2. Spellman, P. T. *et al.* Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3**, RESEARCH0046.1-0046.9 (2002).
3. Brazma, A. *et al.* ArrayExpress — a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71 (2003).
4. Zhao, P., Iezzi, S., Sartorelli, V., Dressman, D. & Hoffman, E. P. Slug is downstream of myoD: identification of novel pathway members via temporal expression profiling. *J. Biol. Chem.* **277**, 20091–20101 (2002).
5. Di Giovanni, S. *et al.* Gene profiling in spinal cord injury shows role of cell cycle in neuronal death. *Ann. Neurol.* **53**, 454–468 (2003).
6. Jin, J. Y., Almon, R. R., DuBois, D. C. & Jusko, W. J. Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays. *J. Pharmacol. Exp. Ther.* **307**, 93–109 (2003).
7. Bakay, M. *et al.* Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformat.* **3**, 4–15 (2002).
8. DePrimo, S. E. *et al.* Expression profiling of blood samples from an SU5416 Phase III metastatic colorectal cancer clinical trial: a novel strategy for biomarker identification. *BMC Cancer* **3**, 3 (2003).
9. de Vos, S. *et al.* Gene expression profile of serial samples of transformed B-cell lymphomas. *Lab. Invest.* **83**, 271–285 (2003).
10. Hittel, D. S., Kraus, W. E. & Hoffman, E. P. Skeletal muscle dictates the fibrinolytic state after exercise training in overweight men with characteristics of metabolic syndrome. *J. Physiol.* **548**, 401–410 (2003).
11. Zambon, A. C. *et al.* Time- and exercise-dependent gene regulation in human skeletal muscle. *Genome Biol.* **4**, R61 (2003).
12. Bakay, M. *et al.* Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformat.* **3**, 4–15 (2002).
13. Cardozo, A. K. *et al.* Gene microarray study corroborates proteomic findings in rodent islet cells. *J. Proteome Res.* **2**, 553–555 (2003).
14. Chun, T. W. *et al.* Gene expression and viral prodution in latently infected, resting CD4+T cells in viremic versus aviremic HIV-infected individuals. *Proc. Natl Acad. Sci. USA* **100**, 1908–1913 (2003).
15. Kamme, F. *et al.* Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. *J. Neurosci.* **23**, 3607–3615 (2003).
16. Huang, J. *et al.* Effects of ischemia on gene expression. *J. Surg. Res.* **99**, 222–227 (2001).
17. Seo, J. *et al.* Interactive color mosaic and dendrogram displays for signal/noise optimization in microarray data analysis. *IEEE ICME* **3**, 461–462 (2003).
18. Somorjai, R. L., Dolenko, B. & Baumgartner, R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* **19**, 1484–1491 (2003).
19. Mei, R. *et al.* Probe selection for high-density oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **100**, 11237–11242 (2003).
20. Li, C. & Hung Wong, W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* **2**, RESEARCH0032 (2001).
21. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
22. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
23. Ambroise, C. & McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA* **99**, 6562–6566 (2002).
24. West, M. *et al.* Predicting the clinical status of human breast cancer utilizing gene expression profiles. *Proc. Natl Acad. Sci. USA* **98**, 11462–11467 (2001).
25. Tusher, V., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5124 (2001).
26. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA* **99**, 6567–6572 (2002).
27. Huang, E. *et al.* Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature Genet.* **34**, 226–230 (2003).
28. Black, E. P. *et al.* Distinct gene expression phenotypes of cells lacking Rb and Rb family members. *Cancer Res.* **63**, 3716–3723 (2003).
29. Huang, E. *et al.* Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596 (2003).
30. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA* **97**, 10101–10106 (2000).
31. Chen, J. *et al.* The PEPR GeneChip data warehouse and implementation of a dynamic time series query tool (SGQT) with graphical interface. *Nucleic Acids Res.* **32**, D578–D581 (2004).
32. Almon *et al. In vivo* multitissue corticosteroid microarray time series available online at Public Expression Profile Resource (PEPR). *Pharmacogenomics* **4**, 791–799 (2003).

Acknowledgements

The authors thank their respective funding agencies, particularly the larger collaborative funding initiatives that make systematic and large-scale studies of the bioinformatics and biostatistics of genome-wide data sets possible from the Department of Defense and the Doris Duke Charitable Foundation CSDA. The authors also thank S. Hilmer, A. DeBlase and G. Miyada for their critique of the manuscript.

Competing interests statement

Some of the authors declare competing financial interests: see Web version for details.

Online links

DATABASES

The following terms in this article are linked online to:
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
GAPDH | *got1*

FURTHER INFORMATION

Affymetrix Developers' Network:
http://www.affymetrix.com/support/developer/index.affx
Affymetrix Technical Note:
http://www.affymetrix.com/support/technical/technotes/blood_technote.pdf
Agilent Technologies: http://www.chem.agilent.com
ArrayExpress microarray database:
http://www.ebi.ac.uk/arrayexpress
The Children's National Medical Center Microarray Center:
http://microarray.cnmcresearch.org/singlegenemain.asp
HOPGENE Program for Genomic Applications:
www.hopkins-genomics.org
MGED Data Transformation and Normalization Working Group: http://www.dnachip.org/mged/normalization.html
MGED Society: http://www.mged.org
NETAFFX web site:
http://www.affymetrix.com/analysis/index.affx
NIGMS Glue Grant:
http://www.gluegrant.org/whatsagluegrant.htm
Access to this interactive links box is free online.

MOLECULAR BIOLOGY OF THE CELL

fourth edition

Bruce Alberts

Alexander Johnson

Julian Lewis

Martin Raff

Keith Roberts

Peter Walter

GS Garland Science
Taylor & Francis Group

Garland
Vice President: Denise Schanck
Managing Editor: Sarah Gibbs
Senior Editorial Assistant: Kirsten Jenner
Managing Production Editor: Emma Hunt
Proofreader and Layout: Emma Hunt
Production Assistant: Angela Bennett
Text Editors: Marjorie Singer Anderson and Betsy Dilernia
Copy Editor: Bruce Goatly
Word Processors: Fran Dependahl, Misty Landers and Carol Winter
Designer: Blink Studio, London
Illustrator: Nigel Orme
Indexer: Janine Ross and Sherry Granum
Manufacturing: Nigel Eyre and Marion Morrow

Cell Biology Interactive
Artistic and Scientific Direction: Peter Walter
Narrated by: Julie Theriot
Production, Design, and Development: Mike Morales

Bruce Alberts received his Ph.D. from Harvard University and is President of the National Academy of Sciences and Professor of Biochemistry and Biophysics at the University of California, San Francisco. **Alexander Johnson** received his Ph.D. from Harvard University and is a Professor of Microbiology and Immunology at the University of California, San Francisco. **Julian Lewis** received his D.Phil. from the University of Oxford and is a Principal Scientist at the Imperial Cancer Research Fund, London. **Martin Raff** received his M.D. from McGill University and is at the Medical Research Council Laboratory for Molecular Cell Biology and Cell Biology Unit and in the Biology Department at University College London. **Keith Roberts** received his Ph.D. from the University of Cambridge and is Associate Research Director at the John Innes Centre, Norwich. **Peter Walter** received his Ph.D. from The Rockefeller University in New York and is Professor and Chairman of the Department of Biochemistry and Biophysics at the University of California, San Francisco, and an Investigator of the Howard Hughes Medical Institute.

Library of Congress Cataloging-in-Publicaton Data
Molecular biology of the cell / Bruce Alberts ... [et al.].-- 4th ed.
p. cm
Includes bibliographical references and index.
ISBN 0-8153-3218-1 (hardbound) -- ISBN 0-8153-4072-9 (pbk.)
1. Cytology. 2. Molecular biology. I. Alberts, Bruce.
[DNLM: 1. Cells. 2. Molecular Biology.]
QH581.2 .M64 2002
571.6--dc21

2001054471 CIP

Published by Garland Science, a member of the Taylor & Francis Group, 29 West 35th Street, New York, NY 10001-2299

Printed in the United States of America

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Front cover Human Genome: Reprinted by permission from *Nature,* International Human Genome Sequencing Consortium, 409:860–921, 2001 © Macmillan Magazines Ltd. Adapted from an image by Francis Collins, NHGRI; Jim Kent, UCSC; Ewan Birney, EBI; and Darryl Leja, NHGRI; showing a portion of Chromosome 1 from the initial sequencing of the human genome.

Back cover In 1967, the British artist Peter Blake created a design classic. Nearly 35 years later Nigel Orme (illustrator), Richard Denyer (photographer), and the authors have together produced an affectionate tribute to Mr Blake's image. With its gallery of icons and influences, its assembly created almost as much complexity, intrigue and mystery as the original. *Drosophila, Arabidopsis,* Dolly and the assembled company tempt you to dip inside where, as in the original, "a splendid time is guaranteed for all."
(Gunter Blobel, courtesy of The Rockefeller University; Marie Curie, Keystone Press Agency Inc; Darwin bust, by permission of the President and Council of the Royal Society; Rosalind Franklin, courtesy of Cold Spring Harbor Laboratory Archives; Dorothy Hodgkin, © The Nobel Foundation, 1964; James Joyce, etching by Peter Blake; Robert Johnson, photo booth self-portrait early 1930s, © 1986 Delta Haze Corporation all rights reserved, used by permission; Albert L. Lehninger, (unidentified photographer) courtesy of The Alan Mason Chesney Medical Archives of The Johns Hopkins Medical Institutions; Linus Pauling, from Ava Helen and Linus Pauling Papers, Special Collections, Oregon State University; Nicholas Poussin, courtesy of ArtToday.com; Barbara McClintock, © David Micklos, 1983; Andrei Sakharov, courtesy of Elena Bonner; Frederick Sanger, © The Nobel Foundation, 1958.)

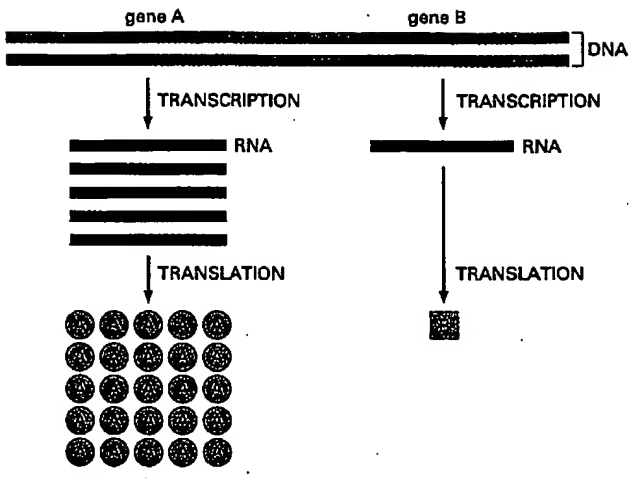

Figure 6–3 Genes can be expressed with different efficiencies. Gene A is transcribed and translated much more efficiently than gene B. This allows the amount of protein A in the cell to be much greater than that of protein B.

FROM DNA TO RNA

Transcription and translation are the means by which cells read out, or express, the genetic instructions in their genes. Because many identical RNA copies can be made from the same gene, and each RNA molecule can direct the synthesis of many identical protein molecules, cells can synthesize a large amount of protein rapidly when necessary. But each gene can also be transcribed and translated with a different efficiency, allowing the cell to make vast quantities of some proteins and tiny quantities of others (Figure 6–3). Moreover, as we see in the next chapter, a cell can change (or regulate) the expression of each of its genes according to the needs of the moment—most obviously by controlling the production of its RNA.

Portions of DNA Sequence Are Transcribed into RNA

The first step a cell takes in reading out a needed part of its genetic instructions is to copy a particular portion of its DNA nucleotide sequence—a gene—into an RNA nucleotide sequence. The information in RNA, although copied into another chemical form, is still written in essentially the same language as it is in DNA—the language of a nucleotide sequence. Hence the name **transcription.**

Like DNA, RNA is a linear polymer made of four different types of nucleotide subunits linked together by phosphodiester bonds (Figure 6–4). It differs from DNA chemically in two respects: (1) the nucleotides in RNA are *ribonucleotides*—that is, they contain the sugar ribose (hence the name *ribo*nucleic acid) rather than deoxyribose; (2) although, like DNA, RNA contains the bases adenine (A), guanine (G), and cytosine (C), it contains the base uracil (U) instead of the thymine (T) in DNA. Since U, like T, can base-pair by hydrogen-bonding with A (Figure 6–5), the complementary base-pairing properties described for DNA in Chapters 4 and 5 apply also to RNA (in RNA, G pairs with C, and A pairs with U). It is not uncommon, however, to find other types of base pairs in RNA: for example, G pairing with U occasionally.

Despite these small chemical differences, DNA and RNA differ quite dramatically in overall structure. Whereas DNA always occurs in cells as a double-stranded helix, RNA is single-stranded. RNA chains therefore fold up into a variety of shapes, just as a polypeptide chain folds up to form the final shape of a protein (Figure 6–6). As we see later in this chapter, the ability to fold into complex three-dimensional shapes allows some RNA molecules to have structural and catalytic functions.

Transcription Produces RNA Complementary to One Strand of DNA

All of the RNA in a cell is made by DNA transcription, a process that has certain similarities to the process of DNA replication discussed in Chapter 5.

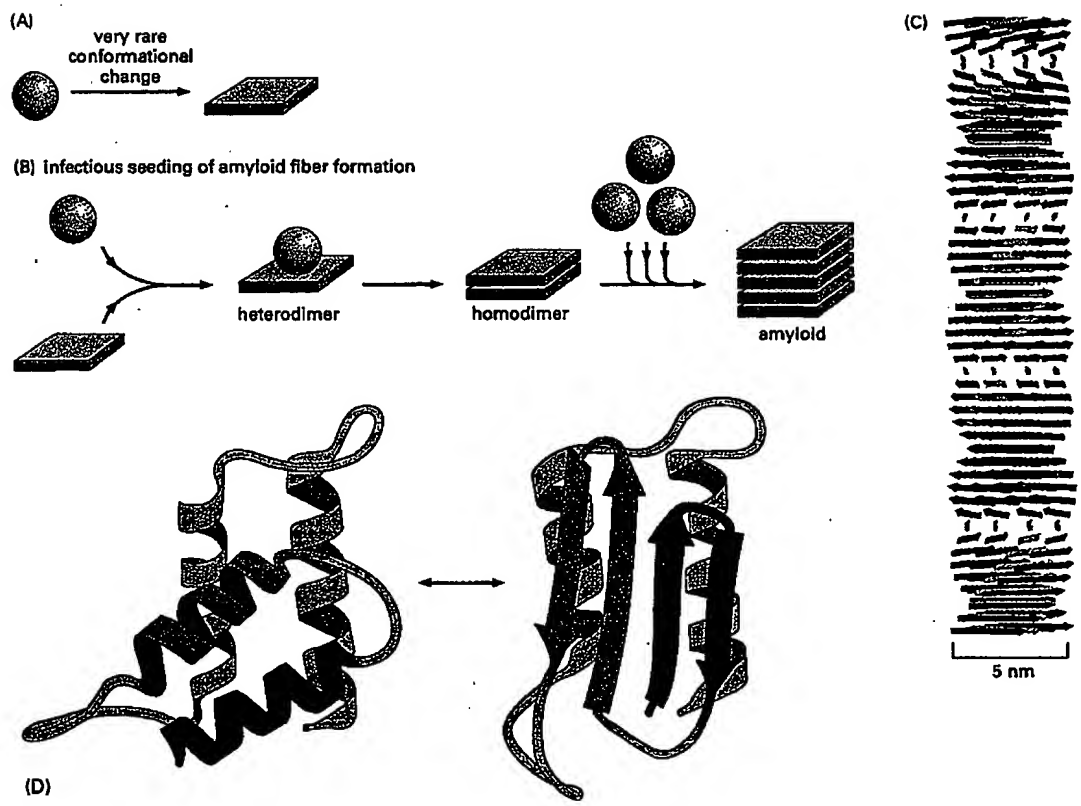

Figure 6–89 Protein aggregates that cause human disease. (A) Schematic illustration of the type of conformational change in a protein that produces material for a cross-beta filament. (B) Diagram illustrating the self-infectious nature of the protein aggregation that is central to prion diseases. PrP is highly unusual because the misfolded version of the protein, called PrP*, induces the normal PrP protein it contacts to change its conformation, as shown. Most of the human diseases caused by protein aggregation are caused by the overproduction of a variant protein that is especially prone to aggregation, but because this structure is not infectious in this way, it cannot spread from one animal to another. (C) Drawing of a cross-beta filament, a common type of protease-resistant protein aggregate found in a variety of human neurological diseases. Because the hydrogen-bond interactions in a β sheet form between polypeptide backbone atoms (see Figure 3–9), a number of different abnormally folded proteins can produce this structure. (D) One of several possible models for the conversion of PrP to PrP*, showing the likely change of two α-helices into four β-strands. Although the structure of the normal protein has been determined accurately, the structure of the infectious form is not yet known with certainty because the aggregation has prevented the use of standard structural techniques. (C, courtesy of Louise Serpell, adapted from M. Sunde et al., *J. Mol. Biol.* 273:729–739, 1997; D, adapted from S.B. Prusiner, *Trends Biochem. Sci.* 21:482–487, 1996.)

animals and humans. It can be dangerous to eat the tissues of animals that contain PrP*, as witnessed most recently by the spread of BSE (commonly referred to as the "mad cow disease") from cattle to humans in Great Britain.

Fortunately, in the absence of PrP*, PrP is extraordinarily difficult to convert to its abnormal form. Although very few proteins have the potential to misfold into an infectious conformation, a similar transformation has been discovered to be the cause of an otherwise mysterious "protein-only inheritance" observed in yeast cells.

There Are Many Steps From DNA to Protein

We have seen so far in this chapter that many different types of chemical reactions are required to produce a properly folded protein from the information contained in a gene (Figure 6–90). The final level of a properly folded protein in a cell therefore depends upon the efficiency with which each of the many steps is performed.

We discuss in Chapter 7 that cells have the ability to change the levels of their proteins according to their needs. In principle, any or all of the steps in Fig-

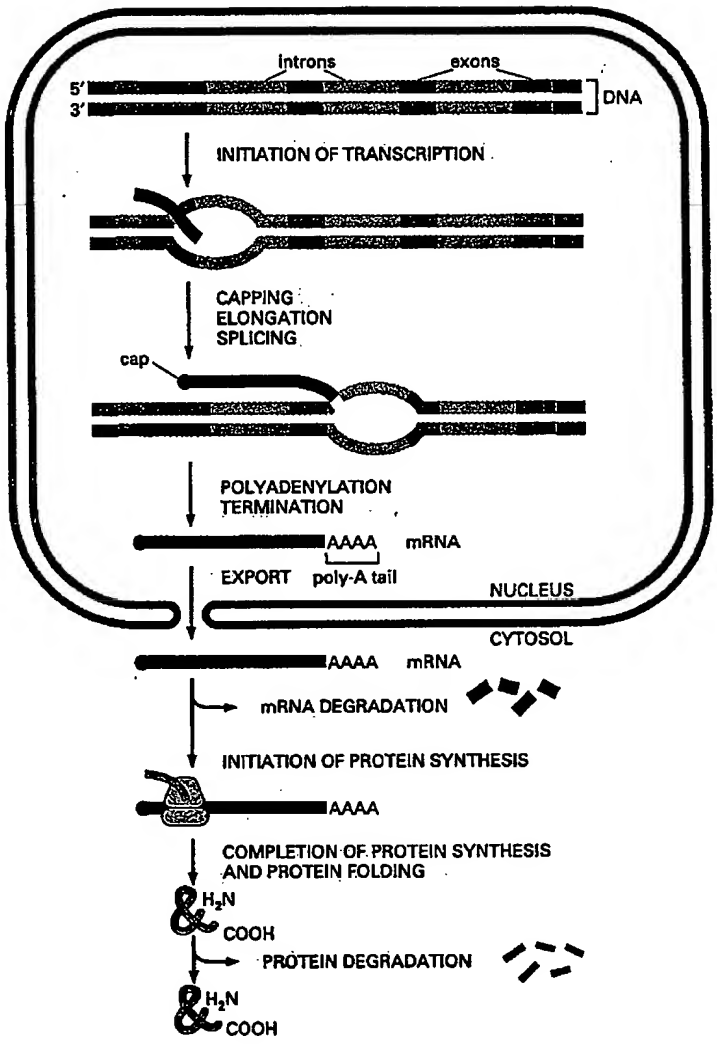

Figure 6–90 The production of a protein by a eucaryotic cell. The final level of each protein in a eucaryotic cell depends upon the efficiency of each step depicted.

ure 6–90) could be regulated by the cell for each individual protein. However, as we shall see in Chapter 7, the initiation of transcription is the most common point for a cell to regulate the expression of each of its genes. This makes sense, inasmuch as the most efficient way to keep a gene from being expressed is to block the very first step—the transcription of its DNA sequence into an RNA molecule.

Summary

The translation of the nucleotide sequence of an mRNA molecule into protein takes place in the cytoplasm on a large ribonucleoprotein assembly called a ribosome. The amino acids used for protein synthesis are first attached to a family of tRNA molecules, each of which recognizes, by complementary base-pair interactions, particular sets of three nucleotides in the mRNA (codons). The sequence of nucleotides in the mRNA is then read from one end to the other in sets of three according to the genetic code.

To initiate translation, a small ribosomal subunit binds to the mRNA molecule at a start codon (AUG) that is recognized by a unique initiator tRNA molecule. A large ribosomal subunit binds to complete the ribosome and begin the elongation phase of protein synthesis. During this phase, aminoacyl tRNAs—each bearing a specific amino acid bind sequentially to the appropriate codon in mRNA by forming complementary base pairs with the tRNA anticodon. Each amino acid is added to the C-terminal end of the growing polypeptide by means of a cycle of three sequential

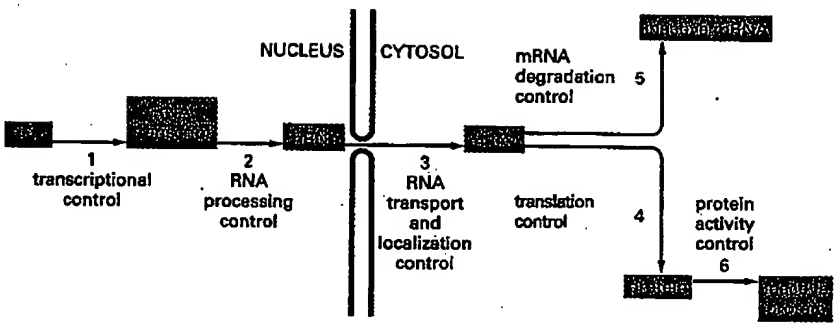

Figure 7–5 Six steps at which eucaryotic gene expression can be controlled. Controls that operate at steps 1 through 5 are discussed in this chapter. Step 6, the regulation of protein activity, includes reversible activation or inactivation by protein phosphorylation (discussed in Chapter 3) as well as irreversible inactivation by proteolytic degradation (discussed in Chapter 6).

Gene Expression Can Be Regulated at Many of the Steps in the Pathway from DNA to RNA to Protein

If differences among the various cell types of an organism depend on the particular genes that the cells express, at what level is the control of gene expression exercised? As we saw in the last chapter, there are many steps in the pathway leading from DNA to protein, and all of them can in principle be regulated. Thus a cell can control the proteins it makes by (1) controlling when and how often a given gene is transcribed (**transcriptional control**), (2) controlling how the RNA transcript is spliced or otherwise processed (**RNA processing control**), (3) selecting which completed mRNAs in the cell nucleus are exported to the cytosol and determining where in the cytosol they are localized (**RNA transport and localization control**), (4) selecting which mRNAs in the cytoplasm are translated by ribosomes (**translational control**), (5) selectively destabilizing certain mRNA molecules in the cytoplasm (**mRNA degradation control**), or (6) selectively activating, inactivating, degrading, or compartmentalizing specific protein molecules after they have been made (**protein activity control**) (Figure 7–5).

For most genes transcriptional controls are paramount. This makes sense because, of all the possible control points illustrated in Figure 7–5, only transcriptional control ensures that the cell will not synthesize superfluous intermediates. In the following sections we discuss the DNA and protein components that perform this function by regulating the initiation of gene transcription. We shall return at the end of the chapter to the additional ways of regulating gene expression.

Summary

The genome of a cell contains in its DNA sequence the information to make many thousands of different protein and RNA molecules. A cell typically expresses only a fraction of its genes, and the different types of cells in multicellular organisms arise because different sets of genes are expressed. Moreover, cells can change the pattern of genes they express in response to changes in their environment, such as signals from other cells. Although all of the steps involved in expressing a gene can in principle be regulated, for most genes the initiation of RNA transcription is the most important point of control.

DNA-BINDING MOTIFS IN GENE REGULATORY PROTEINS

How does a cell determine which of its thousands of genes to transcribe? As mentioned briefly in Chapters 4 and 6, the transcription of each gene is controlled by a regulatory region of DNA relatively near the site where transcription begins. Some regulatory regions are simple and act as switches that are thrown by a single signal. Many others are complex and act as tiny microprocessors, responding to a variety of signals that they interpret and integrate to switch the neighboring gene on or off. Whether complex or simple, these switching devices

occur in the germ line, the cell lineage that gives rise to sperm or eggs. Most of the DNA in vertebrate germ cells is inactive and highly methylated. Over long periods of evolutionary time, the methylated CG sequences in these inactive regions have presumably been lost through spontaneous deamination events that were not properly repaired. However promoters of genes that remain active in the germ cell lineages (including most housekeeping genes) are kept unmethylated, and therefore spontaneous deaminations of Cs that occur within them can be accurately repaired. Such regions are preserved in modern day vertebrate cells as CG islands. In addition, any mutation of a CG sequence in the genome that destroyed the function or regulation of a gene in the adult would be selected against, and some CG islands are simply the result of a higher than normal density of critical CG sequences.

The mammalian genome contains an estimated 20,000 CG islands. Most of the islands mark the 5′ ends of transcription units and thus, presumably, of genes. The presence of CG islands often provides a convenient way of identifying genes in the DNA sequences of vertebrate genomes.

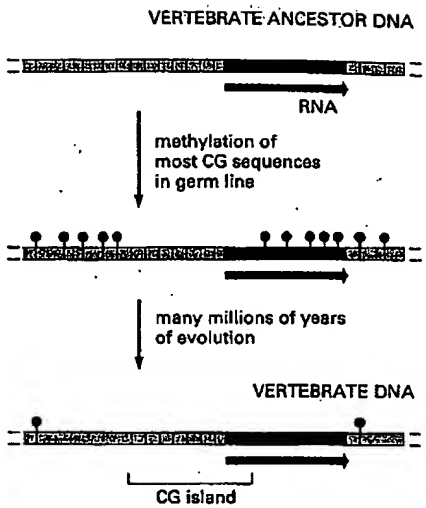

Figure 7–86 A mechanism to explain both the marked overall deficiency of CG sequences and their clustering into CG islands in vertebrate genomes. A *black line* marks the location of a CG dinucleotide in the DNA sequence, while a *red* "lollipop" indicates the presence of a methyl group on the CG dinucleotide. CG sequences that lie in regulatory sequences of genes that are transcribed in germ cells are unmethylated and therefore tend to be retained in evolution. Methylated CG sequences, on the other hand, tend to be lost through deamination of 5-methyl C to T, unless the CG sequence is critical for survival.

Summary

The many types of cells in animals and plants are created largely through mechanisms that cause different genes to be transcribed in different cells. Since many specialized animal cells can maintain their unique character through many cell division cycles and even when grown in culture, the gene regulatory mechanisms involved in creating them must be stable once established and heritable when the cell divides. These features endow the cell with a memory of its developmental history. Bacteria and yeasts provide unusually accessible model systems in which to study gene regulatory mechanisms. One such mechanism involves a competitive interaction between two gene regulatory proteins, each of which inhibits the synthesis of the other; this can create a flip-flop switch that switches a cell between two alternative patterns of gene expression. Direct or indirect positive feedback loops, which enable gene regulatory proteins to perpetuate their own synthesis, provide a general mechanism for cell memory. Negative feedback loops with programmed delays form the basis for cellular clocks.

In eucaryotes the transcription of a gene is generally controlled by combinations of gene regulatory proteins. It is thought that each type of cell in a higher eucaryotic organism contains a specific combination of gene regulatory proteins that ensures the expression of only those genes appropriate to that type of cell. A given gene regulatory protein may be active in a variety of circumstances and typically is involved in the regulation of many genes.

In addition to diffusible gene regulatory proteins, inherited states of chromatin condensation are also used by eucaryotic cells to regulate gene expression. An especially dramatic case is the inactivation of an entire X chromosome in female mammals. In vertebrates DNA methylation also functions in gene regulation, being used mainly as a device to reinforce decisions about gene expression that are made initially by other mechanisms. DNA methylation also underlies the phenomenon of genomic imprinting in mammals, in which the expression of a gene depends on whether it was inherited from the mother or the father.

POSTTRANSCRIPTIONAL CONTROLS

In principle, every step required for the process of gene expression could be controlled. Indeed, one can find examples of each type of regulation, although any one gene is likely to use only a few of them. Controls on the initiation of gene transcription are the predominant form of regulation for most genes. But other controls can act later in the pathway from DNA to protein to modulate the amount of gene product that is made. Although these **posttranscriptional controls**, which operate after RNA polymerase has bound to the gene's promoter and begun RNA synthesis, are less common than *transcriptional control*, for many genes they are crucial.

EXHIBIT 2

REVIEW ARTICLE Medical Biology 62: 304–317, 1984

AMPLIFICATION OF CELLULAR ONCOGENES IN CANCER CELLS

K. ALITALO

FROM THE DEPARTMENT OF VIROLOGY, UNIVERSITY OF HELSINKI, HELSINKI, FINLAND

ABSTRACT

Regulatory or structural alterations of cellular oncogenes have been implicated in the causation of various cancers. Oncogene alteration by point mutations can result in a protein product with strongly enhanced oncogenic potential. Aberrant expression of cellular oncogenes may be due to tumour-specific chromosomal translocations that dysregulate the normal functions of a proto-oncogene. Amplification of cellular oncogenes can also augment their expression by increasing the amount of DNA template available for the production of mRNA. It appears that amplification of certain oncogenes is a common correlate of the progression of some tumours and also occurs as a rare sporadic event affecting various oncogenes in different types of cancer. Amplified copies of oncogenes may or may not be associated with chromosomal abnormalities signifying DNA amplification: double minute chromosomes and homogeneously staining chromosomal regions. Amplified oncogenes, whether sporadic or tumour type-specific, are expressed at elevated levels, in some cases in cells where their diploid forms are normally silent. Increased dosage of an amplified oncogene may contribute to the multistep progression of at least some cancers.

KEY WORDS: CELLULAR ONCOGENES, GENE AMPLIFICATION, MULTISTEP CARCINOGENESIS, CLONAL SELECTION, KARYOTYPIC ABNORMALITIES, DOUBLE MINUTE CHROMOSOMES, HOMOGENEOUSLY STAINING CHROMOSOMAL REGIONS

DNA SEQUENCE AMPLIFICATION AND CYTOGENETIC ABNORMALITIES IN TUMOURS

Since its discovery in drug-resistant eukaryotic cells, somatic amplification of specific genes has been implicated in an increasing variety of adaptive responses of cells to environmental stresses (70, 79). Cytogenetic abnormalities, double minute chromosomes (dmin:s) associated with DNA amplification had already been discovered in tumour cells before the discovery of dmin:s and homogeneously staining chromosomal regions (HSR:s) in cells selected for drug-resistance (12, 24, 49, 50, 56). In metaphase spreads, dmin:s appear as small, spherical, usually paired chromosome — like structures that lack a centromere (Fig. 2). HSR:s stain with intermediate intensity throughout their length rather than with the normal pattern of alternating dark and light bands in both trypsin-Giemsa (Fig. 3A) and quinacrine dihydrochloride-stained preparations. Both kinds of abnormalities are occasionally found in metaphases of freshly isolated cancer cells but not of normal cells (8).

Dmin:s and HSR:s are apparently rare in tumour cells in vivo, although exact data are difficult to obtain since the abnormalities are easily missed in routine cytogenetic analysis (8, 42). Dmin:s and HSR:s have been described in most types of in vitro-cultured malignant tumour cells, with a notable frequency in neuroblastoma cell lines (11). Initial growth in cell culture apparently selects for tumour cells that contain either dmin:s or HSR:s. Moreover, in culture dmin:s are frequently lost, concomitant with the appearance of clonal populations of cells that have developed an HSR, suggesting that the two cytogenetic abnormaltities are alternative forms of gene amplification and that HSR:s may confer a selective advantage on cells over dmin:s (11, 70). It has been assumed that HSR:s can break down to firm dmin:s and that dmin:s can integrate into chromosomes to generate HSR:s (11, 23). Amplified genes may also occupy abnormally banding regions, ABR:s (51, 59). Experimental work on drug-resistant cells has shown that in the absence of a selection pressure (drug), dmin:s and the amplified genes that they contain are lost, whereas amplified DNA in the form of HSR:s is retained in the cells (71). This is explained by the fact that dmin:s are segregated unevenly in mitosis and frequently get lost from the nucleus due to

TABLE 1

Currently known oncogenes.

Retrovirus (example)	Oncogene	Gene product		
		Cellular location	Function of protein	Class
ONCOGENES FOUND IN RETROVIRUSES				
RSV	*src*	Plasma membrane	Tyrosine-specific protein kinases (*fgr* contains sequences homologous to actin)	Class Ia (Cytoplasmic tyrosine protein kinases)
Y73V	*yes*	Plasma membrane		
GR-FeSV	*fgr*	Plasma membrane		
Ab-MuLV	*abl*	Plasma membrane		
FuSV	*fps/fes*	Cytoplasm (plasma membrane?)		
ST-and GA-FeSV	*fes/fps*	Cytoplasm (cytoskeleton?)		
UR2V	*ros*			
AEV	*erb*-B	Plasma membrane and cytoplasmic membranes	EGF receptor's cytoplasmic domain	Class 1b (Class 1a-related proteins)
SM-FeSV	*fms*	Plasma membrane and cytoplasmic membranes	Cytoplasmic domain of a growth factor receptor?	
MH-2V	*mil/raf*	Cytoplasm	?	
3911-MSV	*raf/mil*	Cytoplasm	?	
Mo-MSV	*mos*	Cytoplasm	?	
SSV	*sis*	Secreted	PDGF-like growth factor	Class 2 (Growth factors)
Ha-MSV	Ha-*ras*	Plasma membrane	GTP-binding proteins	Class 3 (Cytoplasmic GTP:ases)
Ki-MSV	Ki-*ras*	Plasma membrane		
FBJ-MuSV	*fos*	Nucleus	?	Class 4 (Nuclear phosphoproteins)
OK-10V	*myc*	Nucleus	Nuclear matrix protein	
AMV	*myb*	Nucleus	?	
SKV 770	*ski*	Nucleus?	?	
REV	*rel*	?	?	Unclassified
AEV	*erb*-A	?	?	
E26V	*ets*	?	?	
ONCOGENES FOUND IN TUMOUR CELLS BUT NOT IN RETROVIRUSES				
Tumour cell				
Neuroblastoma	N-*ras*	Plasma membrane	GTP—binding	Class 3
Neuroblastoma	N-*myc*	?	?	Class 4
Small-cell lung cancer	L-*myc*	?	?	Class 4
Neuro-/Glioblastomas	*neu*	Plasma membrane	Growth factor receptor	Class 1b

their lack of centromeres, (49). HSR chromosomes carry centromeres and are therefore divided equally at mitosis. If dmin:s and HSR:s contain amplified genes that encode growth-stimulating protein products, it would follow that the more stable chromosomal form, the HSR, confers a greater selective growth advantage for cells. Although dmin:s and HSR:s have been described predominantly in tumour cells selected for resistance to cytotoxic drugs, it is also clear that dmin:s and HSR:s may be present in cancer cells before any form of therapy (8). It was in this setting that we and others first chose to explore the possible amplification of cellular oncogenes. (By definition, cellular oncogenes are normally innocent genetic loci which can be activated to transforming genes in various ways).[1]

DMIN:S AND HSR:S CONTAIN AMPLIFIED ONCOGENES

Table 2 summarizes the somatic amplifications of cellular oncogenes so far reported in

[1] It is not the purpose of this review to deal with all forms of DNA damage that have been found to activate cellular oncogenes. For the purpose of integrating the review into a coherent picture, however, the reader is given a list of known cellular oncogenes in Table 1 and the schematic Figure 1 illustrating the various ways in which the oncogenic potential of different proto-oncogenes can be activated.

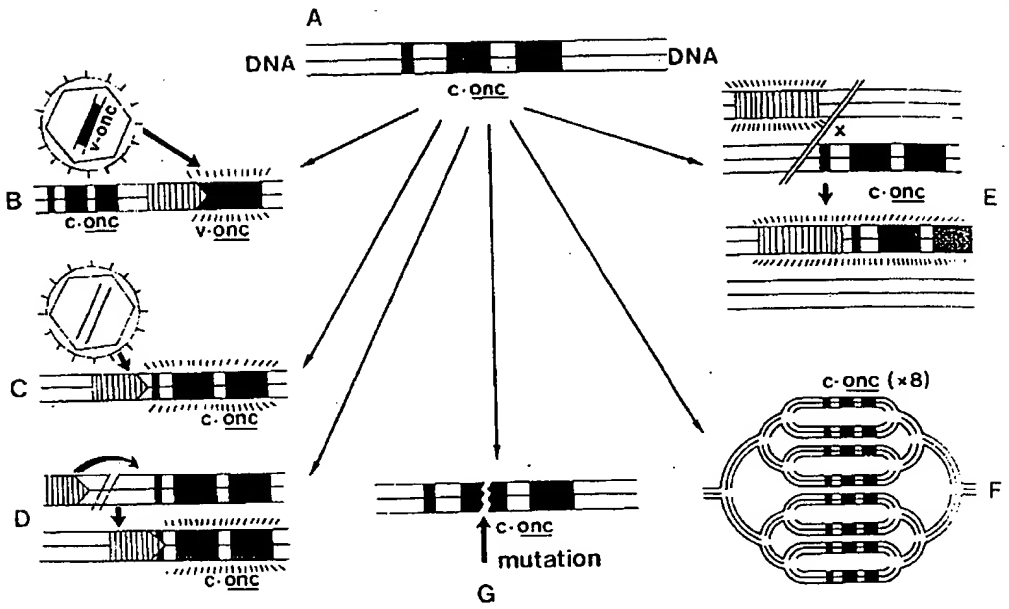

Fig. 1. Activation of cellular oncogenes. The haploid complement of a proto-oncogene is schematically depicted in A, composed of three exons (black boxes) in a segment of DNA. The different activated forms are schematically outlined in B—G. The abbreviation c-*onc* stands for cellular oncogene, v-*onc* viral oncogene, DNA sequences with associated strong promoter/enhancer functions are striated, and an actively transcribed gene is marked with radiations. B. Acute transforming retroviruses have the capacity to transduce cellular oncogenes (c-*onc*) into their genome, modify them and reinsert their activated oncogenes (v-*onc*) into the genome of host animal cells as a part of their proviral forms. The activity of the v-*onc* gene is greatly enhanced due to the associated promoter of the proviral long terminal repeat. Both increased dosage of the oncogene and structural mutations within its sequence may contribute to tumorigenesis. C. Slow transforming retroviruses without oncogenes replicate and reinsert their proviral copies into the host cell DNA during a latency period from infection to tumorigenesis. Tumor initiation through hyperplastic growth may begin, when the provirus integrates sufficiently close to a proto-oncogene to activate it through promoter or enchancer functions. It should be noted, however, that mutations have also been found in the oncogenes thus activated and that mutational damage to other oncogenes has been described in the resulting tumors. D. In some mouse plasmacytomas, a retrovirus-like DNA element (directing the synthesis of the so-called intracisternal A-type particles, IAPs) has been found in association with a transcriptionally activated oncogene c-*mos*. The IAP insertion also disrupts the 5' part of c-*mos* (64). E. In humans, as well as in animals, chromosome translocations may place proto-oncogenes into transcriptionally active regions of chromatin, where they may be activated. The details of this mechanism have not been worked out, but it is believed to occur for c-*myc* and c-*abl* genes in Burkitt lymphomas and Philadelphia-chromosome positive leukemias, respectively (35, 40). F. Increased amounts of oncogene-specific RNA and protein can also result from an excess of DNA template for transcription acquired through oncogene amplification. The present review concentrates primarily on this mechanism. G. Mutationally activated oncogenes have been found in nearly one fifth of human malignant tumours. Oncogene loci activated by somatic structural mutations are revealed by transfection experiments, where they are introduced into genetic background of nontumourigenic cultured immortalized cells. Several such transforming loci have been cloned and many of them belong to the c-*ras* oncogene family. It should be pointed out that both structural mutations and either increased expression or activation of a complementing oncogene may be required to achieve a fully tumorigenic phenotype (44).

tumour cells. Although the sampling of tumours is at present small, the finding of known cellular oncogenes among amplified DNA represented by dmin:s and HSR:s of cancer cells is provocative. Amplification has been found to affect at least five out of twenty known cellular oncogenes and the degree of gene amplification varies from five to many hundred-fold over the single haploid copies found in normal cells (see also ref. 18). The first amplification reported involved the c-*myc* oncogene (see Table 1) in a promyelocytic leukaemia cell line HL-60 (20, 25). The degree of c-*myc* amplification is between 8—32 fold both in the HL-60 cell line and in primary leukaemic cells from the patient (20, 25). Original clonal lines of HL-60 were later found to contain some dmin:s in culture but their number was insufficient to establish any clear correlation with amplified c-*myc*. Such a correlation, however, was discovered for c-*myc* amplification in a neuroendocrine cell line from

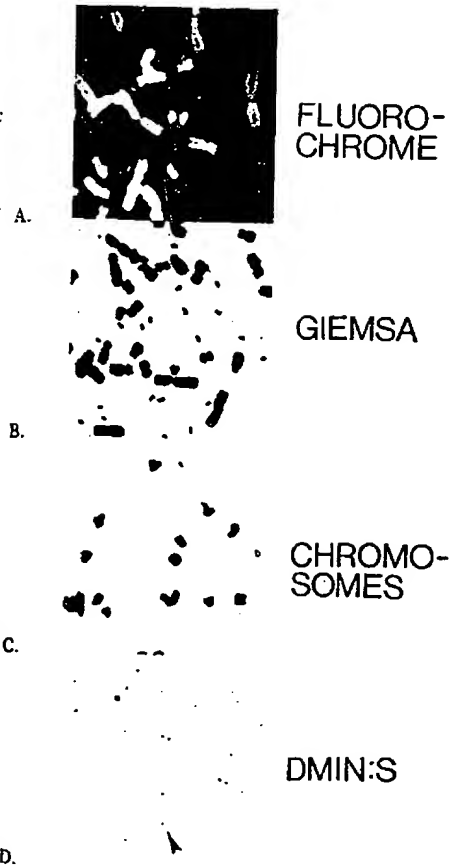

Fig. 2. Double minute chromosomes (arrowheads) in COLO 320DM colon carcinoma cells. A. The dmin:s are resolved as paired dots among normal chromosomes in this fluorescent, benzimidazole-stained preparation B—D. Purification of dmin:s by differential centrifugations. B. The starting material, C. Chromosome fraction, D. Purified dmin:s (Donna George and the author, unpublished data and ref. 52).

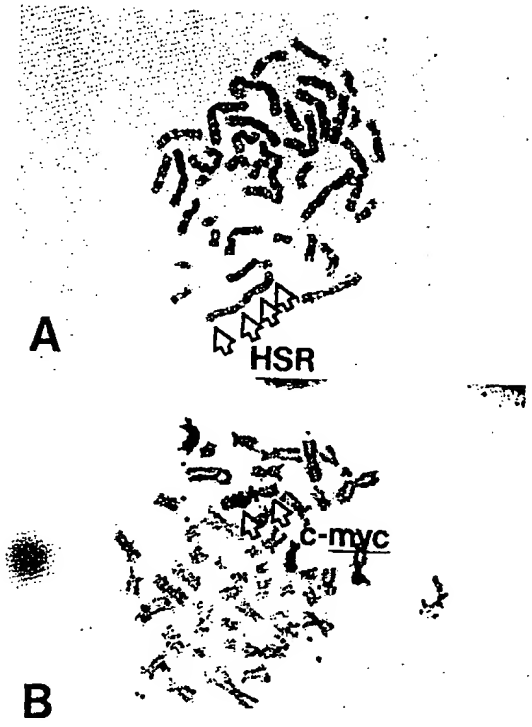

Fig. 3. A. The homogeneously staining regions (HSR) in the G-banded HSR-marker chromosome comprise a major portion of both its long and short arms. The HSR-marker chromosome has evolved from an X-chromosome (52 and unpublished data of C. C. Lin and the author). B. The about 30-fold amplified copies of the c-*myc* oncogene in COLO 320 cells were found to be located to dmin:s and HSR:s. The latter is shown here by in situ-hybridization (5, 52).

a colon carcinoma, COLO 320 (5). In these cells, the approximately 30-fold amplified c-*myc* copies were mapped either to HSR:s of a marker chromosome (5, Fig 3B) or to dmin:s (52), depending on the particular subline studied. Since dmin:s were already present in the primary tumour cells from this colon carcinoma (63), it is likely that c-*myc* had also been amplified during growth of the tumour in vivo. Similarly, amplified copies of the c-Ki-*ras* oncogene were mapped to dmin:s and HSR:s of a mouse adrenocortical tumor Y1 (74). An extensive search for changes in other oncogenes and tumour cells has since revealed amplifications that do not show up as dmin:s or HSR:s. Thus, for example, the c-*myb* oncogene is amplified in a characteristic marker chromosome of a colon carcinoma without evidence of HSR:s (ref. 6, Fig. 4) and in other tumours, the amplified c-*abl* and c-*myc* oncogene loci map to abnormally banding regions (ABR:s) in translocated or resident chromosomal segments, respectively (59, 76).

TRANSLOCATIONS AND REARRANGEMENTS MAY ACCOMPANY ONCOGENE AMPLIFICATION

The evolution and progression of the karyotype of tumour cells is complicated (see ref. 68). Concomitant with amplification, DNA sequences acquire an increased mobility in the genome with extrachromosomal intermediates

TABLE 2

Sporadic and tumour-specific amplification of cellular oncogenes.[*]

Tumour cells	Onco-gene	Fold amplified	Chromosomal location of amplified gene	Expression elevated	Remarks	References
Sporadic:						
HL60 (acute promyelocytic leukaemia, M3)	c-*myc*	20×	8q(ABR)	Yes	Amplification present in primary leukaemic cells	20, 25, 59
COLO320 (colon carcinoma)	c-*myc*	30×	dmin, HSR	Yes	Part of the amplified c-*myc* sequences rearranged	4, 5, 52
Y1 (adrenocortical tumour)	c-Ki-*ras*	50×	dmin, HSR	Yes	Levels of p $21^{c\text{-}Ki\text{-}ras}$ protein elevated	74
COLO201/205 (colon carcinoma)	c-*myb*	10×	mar1	Yes	Patient treated with 5-fluorouracil prior to culturing of the tumour cells	4, 6,88
K562 (chronic myelogenous leukaemia, CML)	c-*abl*	10×	mar(ABR)	Yes	C_λ coamplified in the marker that may be derived from chromosome 22, c-*abl* protein-associated tyrosine kinase activated	21, 22, 41, 54, 76
A431 (epidermoid carcinoma)	c-*erb*B	15—20×	n.d.	Yes	Amplification linked to chromosome 7 translocation and sequence rearrangements Amount of protein product, the EGF receptor, elevated	82 (see 36)
ML1—3 (acute myeloid leukaemia, M2)	c-*myb*	5—10×	n.d.	Yes	Abnormalities of chromosome 6q22—24, where c-*myb* is normally located	34, 61, 91
SK BR-3 (breast carcinoma)	c-*myc*	10×	n.d.	Yes		43
SEWA (polyoma virus-induced mouse tumour)	c-*myc*	30×	n.d.	Yes	Cells have dmin:s depending on culture conditions; c-*myc* amplification correlates with growth as a tumour	Manfred Schwab, personal communication
Lu-65 (lung giant cell carcinoma)	c-*myc* c-Ki-*ras*	8× 10×	n.d. n.d.	n.d. n.d.	At least some copies of c-Ki-*ras* mutated	80
Primary leukemic cells from an acute myeloid leukemia (M2) patient	c-*myc*	33×	n.d.	n.d.		Unpublished data of the author and A. de la Chapelle
Tumour-specific:						
small-cell lung cancer	c-*myc* L-*myc*, N-*myc*	up to 80×	n.d.	Yes	Most amplifications in the variant phenotype of SCLC	53, 69
Neuroblastomas	*N-myc*	up to 250×	dmin, HSR	Yes	N-*myc* also amplified in primary tumours of advanced grade	14, 48, 72, 73, 75
Glioblastomas	c-*erb*B	—	—	—	Rearrangements also found personal communication	Josef Schlessinger, personal communication

n.d. = not determined, mar = marker chromosome, M2, M3 refer to the French-American-British classification of acute myeloid leukemias.

[*] At least one case of oncogene amplification in normal germ-line cells has been found (18).

visualized as dmin:s, transpositions and translocations to other chromosomal segments, etc. (see 70 for references). There may not be preferred chromosomal sites for the apparent reintegration of dmin:s as HSR:s (75). In at least one case, however, an oncogene may have been caught amplifying in situ in its resident chromosomal site (59). The finding of moderately amplified oncogenes also in chromosomal sites lacking HSR:s suggests that

Fig. 4. Localization of amplified c-*myb* in COLO 201/205 cells by in situ hybridization. Shown is a characteristic, large marker chromosome (mar1) with G-banding (A) and associated c-*myb* autoradiographic grains (B). Note the absence of HSR:s. Mar1 has probably evolved from chromosome number 6, the resident site of the c-*myb* oncogene in normal cells (34, 88, 91). (Robert Winqvist and the author, unpublished data).

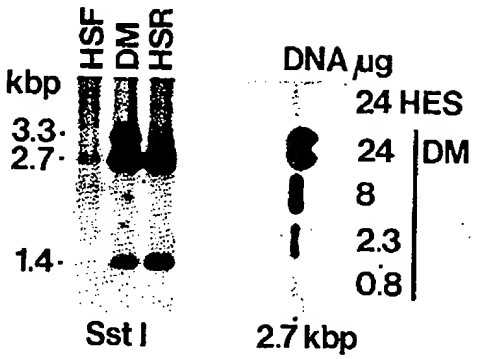

Fig. 5. Amplification and rearrangement of c-*myc* in COLO 320 cells. 10 µg of cellular DNA was digested with Sst I, electrophoresed, blotted and probed with a v-*myc* Pst I fragment (ref. 2, left panel). Fragments of 2.7 kbp and 1.4 kbp are seen in both normal and amplified c-*myc* DNA. The 3.3 kbp fragment is derived from a DNA segment of unknown origin translocated to the 5' region of c-*myc* with a concomitant deletion of its first exon (unpublished data of Manfred Schwab and the author). HSF, human skin fibroblasts; DM, COLO 320 DM cells; HSR, COLO 320 HSR cells. Different amounts of DNA from COLO 320 DM cells as indicated were mixed with calf thymus DNA to give 24 µg of total DNA, cleaved with Sst I, electrophoresed, blotted and probed with a fragment of 3'human c-*myc* sequences. The intensities of the 2.7 kbp c-*myc* fragment in different samples were compared to assess its copy number, estimated to be about 30 (5).

(onco)gene amplification may be more common than the structural alterations shown by chromosome banding and microscopy (6, 88).

In at least three cases reported amplification has been accompanied by a DNA rearrangement of the oncogene (5, 20, 82). In the colon carcinoma COLO 320 both damaged and normal versions of the c-*myc* gene are amplified (5). Although individual cell clones have not yet been examined, our unpublished experiments suggest that the same dmin-containing cells harbor and express both normal and rearranged forms of c-*myc*. The normal version of the amplified gene, however, predominates in COLO 320 cells containing HSR:s; the rearranged version is present only in what appears to be a single copy (Fig. 5). In the chronic myeloid leukaemia (erythroleukaemia) cell line K562 an amplified DNA segment consists of portions of both the c-*abl* oncogene and the immunoglobulin C_λ locus (76). In both cases abnormal transcripts are produced from the rearranged amplified oncogenes (Fig. 6 and ref. 22). In K562 cells, the abnormal c-*abl* oncogene product has also been activated as a tyrosine protein kinase (41). It is not known, however, whether structural alterations of the genes preceded amplification or whether they were acquired during the process of gene amplification. It seems likely that a chromosomal translocation of c-*abl* to the C_λ locus preceded DNA amplification in the K562 cells, since all amplified copies were also rearranged (21), with the change reminiscent of the Philadelphia translocation (t(9, 22)) found in most chronic myeloid leukaemia tumours (35, 66—68). Although they have not been sequenced, other reported cases of amplified oncogenes are apparently normal on basis of mapping with restriction endonucleases (see Table 2). Therefore we cannot at present view mutation as a necessary companion of oncogene amplification.

THE MECHANISMS OF GENE AMPLIFICATION

The mechanisms of gene amplification and the structure of the amplified DNA have been worked out mainly in experimental settings involving selection for drug-resistance in cell culture (70). Although the mechanisms are still incompletely known and may vary in different cases, some general features have emerged.

A spontaneous degree of illegitimate DNA replication seems to exist in normal cells so that various segments of DNA are replicated more than once during a single cell cycle (37). In unselective conditions this DNA is probably lost e.g., through formation of micronucleae

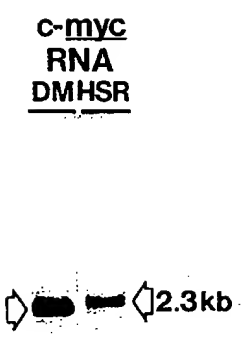

Fig. 6. Comparison of the electrophoretic mobilities of c-*myc* mRNA:s from COLO 320 DM and HSR cells. The size of the normal c-*myc* mRNA is 2.3 kb. The rearranged c-*myc* locus in DM cells (see Fig. 5) seems to be predominantly expressed giving rise to a shortened RNA.

because the newly synthesised extra copies of DNA are not covalently linked to chromosomal DNA of mitotic cells (65, 71). If, however, there is a selection pressure to retain an increased gene dosage, progressive multiplication of gene copy number results. The incidence of cells bearing amplified genes under conditions of cytotoxic selection can vary by two orders of magnitude and is greatly increased by the presence of mitogenic substances (hormones or tumour promoters) during selection (10, 84, 85) or certain carcinogenic or cytotoxic agents before selection (15, 55, 79, 80, 81, 85). An interesting hypothesis suggested by Varshavsky (84, 85) supposes that the origins of DNA replication "fire" (initiate replication) illegitimately several times during a single cell cycle and that this kind of "replicon misfiring" may be increased by substances such as tumour promoters and mitogenic hormones (10, 84, 85). Mariani and Schimke (55) point out that most of the cytotoxic agents that increase the incidence of gene amplification are inhibitors of DNA synthesis. Aberrant replication is known to take place after transient inhibition of DNA synthesis and this response can lead to gene amplification (46, 47, 55, 90). Mitogenic hormones probably increase disproportionate DNA replication, but they also enhance the colony forming efficiency of drug-resistant cells in selective conditions (10).

According to the studies of Axel and his collaborators (63), the multiple cycles of unscheduled DNA replication at a single locus during a single cell cycle result in a structure schematically outlined in Fig. 1F. The hydrogen-bonded amplified copies of DNA depicted in Fig. 1F must resolve into a tandem linear array before the next mitosis. This may well occur by homologous recombination between any one of several repeated sequences within the amplified domain (45, 65). Part of the recombinations would lead to extrachromosomal circles possessing an origin for replication (16, 62); these could be the precursors of dmin:s. The unequal recombinations mean that the resolved linear structure consists of tandemly repeated but heterogeneous units. According to Axel's model a gradient of amplification is formed so that centrally located sequences are amplified more than sequences distal to the origin of replication (65). This also has, in fact, been found to explain the large, complex DNA domain amplified in neuroblastoma cells in vivo (38, see also below).

The chromosomal site of integration of transfected genes significantly affects the frequency and cytogenetic result of their experimentally induced amplification (83). The amplification frequency in some transformants has been found to be 100-fold that of the others (83). This suggests that there also are preferred chromosomal positions for amplification of host cellular genes and that chromosomal rearrangements may facilitate gene amplification by positioning chromosomal sequences in a favorable array. In respect of the structural properties of the sequences involved in gene amplification, recombinatorially active regions have been implicated in experimental cases. DNA rearrangements involving restriction fragment length polymorphisms and variation in gene copy number have been detected in the human genome between clusters of short repetitive interspersed DNA sequences called Alu family DNA-sequences (17). Such inter-Alu sequences have also been detected in an extrachromosomal DNA form, including covalently closed circles (17, 78). The copy number of inter-Alu sequences apparently varies in an age- and tissue-specific manner (17, 78), but any comprehensive analysis of the phenomenon in human tumours is not yet available. It is also not yet clear whether these kinds of repetitive sequences are involved in generating amplified oncogene sequences in dmin:s or HSR:s in tumours.

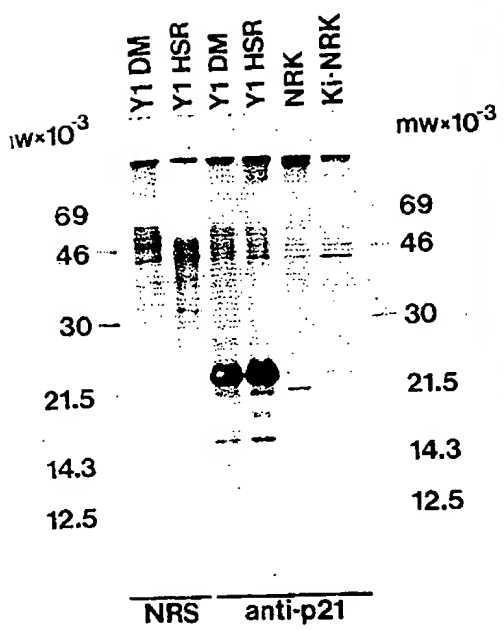

Fig. 7. Elevated levels of the p21$^{c\text{-}Ki\text{-}ras}$ protein in Y1 cells (74). The Y1 DM and HSR cells which harbor a 50-fold amplified c-Ki-*ras* oncogene (74) and control cells were labeled with [^{35}S]-methionine and the p21$^{c\text{-}Ki\text{-}ras}$ protein was immunoprecipitated, as detailed (74), with normal rat serum (NRS) or rat monoclonal anti-p21 serum. The proteins were electrophoresed in a 15 % polyacrylamide gel in the presence of SDS. In addition to a major p21 band, a labeled band at about 16 kd was present in the immunoprecipitates. The amount of radioactivity in p21 was about 50 fold that in normal rat kidney cells. The Kristen sarcoma virus-transformed rat kidney cells (obtained from the American Tissue Culture Collection) also yielded unexpectedly low amounts of the v-Ki-*ras* protein.

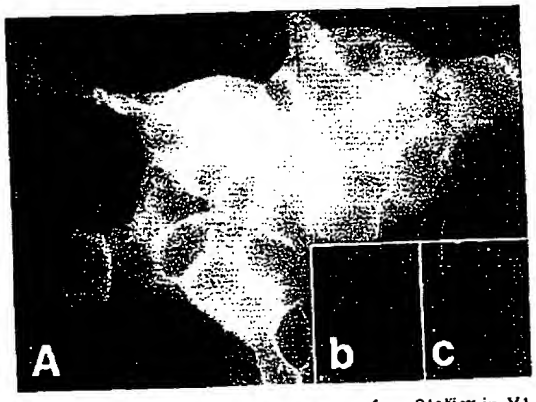

Fig. 8. A. Indirect immunofluorescence for p21$^{c\text{-}Ki\text{-}ras}$ in Y1 DM cells. Similar fluorescence of the plasma membrane was obtained for the Y1 HSR cells. Inset (b) shows control staining with normal rat serum and inset (c) staining of normal rat kidney cells with the monoclonal antibody against p21.

CARCINOGEN-INDUCED GENE AMPLIFICATION AND CLONAL SELECTION OF CANCER CELLS

Although cell sorting experiments have shown a basal spontaneous rate of gene amplification in eukaryotic cells (37), this can be increased severalfold by metabolic inhibitors or cytotoxic agents (15, 37, 70, 81, 85). In many respects the latter response is reminiscent of the so-called SOS-response elicited in bacteriae by noxious stimuli (see 28). In a teleological context, the rapid induction of gene amplification that apparently occurs frequently through extrachromosomal intermediates may provide cells with genetic material for subsequent selective pressures operating in harmful conditions (60). In cancer cells, the mechanism may enhance the emergence of clonal populations of cells with increasingly malignant properties (58, 60). Such genetic instability of cancer cells is clearly enhanced, leading to the rapid evolution of increasingly malignant tumour cell populations (19, 58). A serious question of practical importance is whether drug resistance in treated patients also selects cells that have an enhanced ability to amplify (onco)genes important for growth and progression of the tumour (84, 85). It is also possible that some of the carcinogenic insults caused by mutagens are only expressed as a result of subsequent amplification events induced by tumour promoters (84, 85) or facilitated by hormones in, say replicating epithelial cells (10). The persistence of dmin:s in some tumours suggests that there is a selection pressure for their retention (8, 9, 11, 23). Amplified DNA in dmin:s must contain an origin for DNA replication (62) and must be selected for in daughter cell populations, where it is unevenly segregated (71). In the absence of such a selection pressure dmin:s are lost (71). In at least one study the length of an HSR has been found to increase during a selection of malignant cells for enhanced tumourigenicity (30).

The amplified c-*erb*B gene in A431 cells codes for epidermal growth factor receptor (27). The abundant amounts of receptor protein on A431 cell surface may, however, provide the cells with an abnormal growth response (31). A naive supposition is that the amplified sequences in dmin:s and possibly in HSR:s of tumours contain growth-promoting genes (see 36 for references). This seems to fit well with

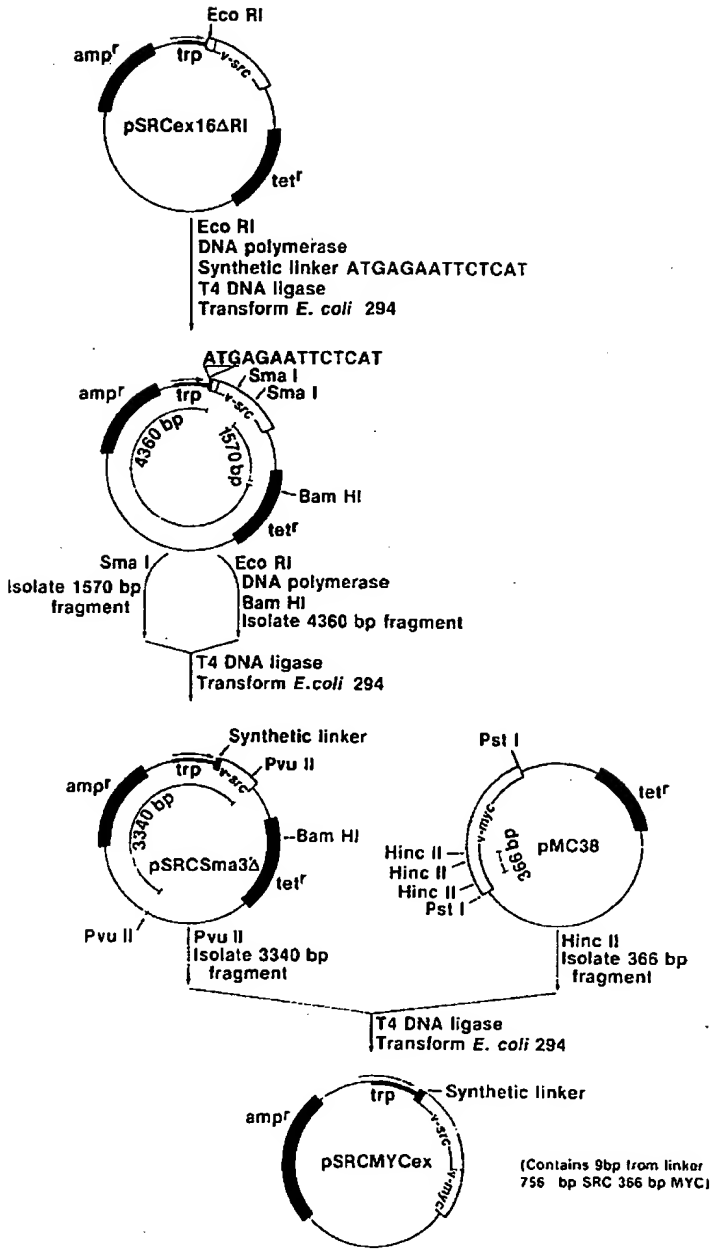

Fig. 9. Construction of a v-*myc* expression vector. A synthetic linker (ATGAGAATTCATCAT) containing a translational initiation codon was inserted downstream from the *trp* promoter in the pSRC ex16 RI expression vector described previously (see ref. 3). Approximately one half of the v-*src* sequences coding for the aminoterminal portion of pp60$^{v\text{-}src}$ protein were then deleted and the remaining portion ligated in translational codon frame with the synthetic ATG. A Hinc II fragment of v-*myc* from plasmid clone MC 38 (nucleotides 320—685 in the v-*myc* sequence in ref. 2) was ligated downstream form remaining v-*src* sequences in continuity with its reading frame. The resulting product contained 3 amino acids from the synthetic linker, 252 amino acids encoded by the 756 base pair fragment from Sma I to Pvu II restriction sites in v-*src* DNA, 122 amino acids from the v-*myc* and 6 amino acids (corresponding to nucleotides 2968—2085) from the pBR322 vector (3).

recent findings on amplified oncogenes, though in many cases the search for an amplified oncogene is still continuing. Even positive findings do not mandate a role for amplified cellular oncogenes, however, because the domain of amplified DNA is inevitably much larger than a single genetic locus (e.g. 38).

ENHANCED EXPRESSION OF AMPLIFIED ONCOGENES

In all cases where they have been studied, the amplified oncogenes have been found abundantly expressed at the RNA level, roughly in proportion to the amount of DNA amplification (see Table 1). Described cases of elevated RNA expression include examples of abnormal (5, 22) and ectopic (6) transcription. In at least four cases this enhancement is not limited to synthesis of RNA (31, 33, 41, 74, 82). The Y1 cells that have amplified c-Ki-*ras* contain exceptionally large amounts of its protein product situated on the plasma membrane (ref. 74, Fig. 7 and 8). High amounts of the c-*myc* encoded protein are also found in COLO 320 cells that have amplified the gene (33). The *myc* oncogenes have recently been shown to encode nuclear proteins (ref. 1, 3, 26, 29, 32, 33, Fig. 9—11). Both the expression of the c-*myc* mRNA (39) and the subcellular localization of myc proteins are linked to the cell cycle (ref. 89, Fig. 12). It may be that elevated expression of specific c-*myc* functions is necessary for cell cycle progression and the growth transformation aspect of the phenotype of cancer cells that may contribute to tumour progression (7, 36). Elevated expression of c-*myc* has been shown to replace in part platelet-derived growth factor in induction of competence for DNA replication (7). Generally, enhanced expression of an oncogene could be a necessary prerequisite for acquisition of a growth advantage by cells having extra copies of the gene. This effect could also be the principal contribution of amplification to tumourigenesis.

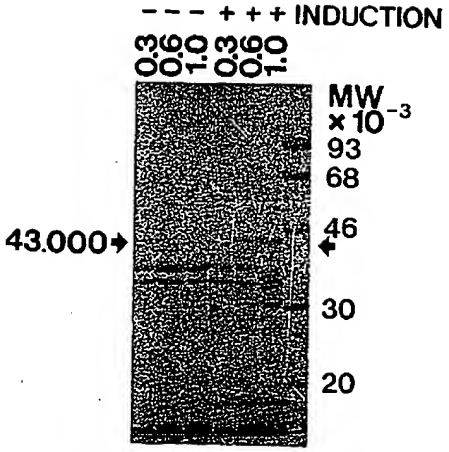

Fig. 10. *E. coli* 294 was transfected with the hybrid v-*src*, v-*myc* plasmid outlined in Fig. 9 and ampicillin-resistant bacterial colonies were checked for the production of a 43,000 m.w. bacterial v-*myc* protein after induction by growth to different optical densities in minimal essential medium (M9, induction +) or complete medium (LB, induction—) (3).

TUMOUR CELL AND STAGE SPECIFICITY OF ONCOGENE ACTIVATION AND AMPLIFICATION

Tumour cell specificity of oncogene amplification has been found in three malignancies. The c-*myc*, L-*myc* or N-*myc* oncogene is amplified in most cases of the variant form of small-cell lung cancer cells (53, 69), c-*erb*B is amplified in several glioblastomas (Josef Schlessinger, personal communication) and the putative N-*myc* oncogene is amplified in about half of grade III and IV neuroblastomas (14, 72, 73, 75). In addition to HSR:s, small-cell lung cancers and neuroblastomas frequently show a

Fig. 11. Indirect immunofluorescence for the v-*myc* protein and phase contrast microscopy of myelocytomatosis virus-transformed quail cells (3). A. Quail cells transformed with the MC-29 virus (Q8 cells). Anti-*myc* protein staining. B. Phase contrast microscopy of field in A. C. Q8 cells stained with anti-*myc* protein antiserum that has been blocked with the immunogen. D. Phase contrast microscopy of field in B. E. Q8 cell stained with preimmune rabbit serum. F. Phase contrast microscopy of field in E.

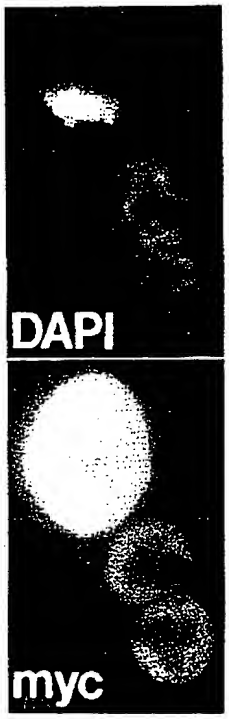

Fig. 12. Fluorescent staining for DNA and *myc* protein in myelocytomatosis virus-transformed quail cells. In interphase cells, the *myc* protein is confined to the nucleae. In the mitotic cell, *myc* fluorescence is distributed throughout the cell unlike fluorescence for chromatin, which is compacted to chromosomes in the metaphase plate. In fact, there is less *myc* fluorescence associated with chromatin than with the rest of the cell. DAPI, diaminophenylindole DNA stain. The anti-*myc* protein rabbit antiserum was used in a 1/200 dilution (ref. 89).

deletion of a portion of the short arm of chromosome 1 (13) and chromosome 3 (86, 87), respectively, in karyological examination. Two kinds of changes have also been described in different neuroblastoma oncogenes. The first is a mutation in the N-*ras* gene, an activated oncogene that was discovered because of its relation to other *ras* genes and transforming activity in transfection experiments (77). The second is amplification of a distant homologue of the c-*myc* gene called N-*myc* (72, 73, 75). Although the transforming potential of the N-*myc* gene has not yet been established, its consistent presence in a core segment of amplified neuroblastoma DNA (38, 57, 72, 73, 75) and its elevated expression in most retinoblastomas (48) suggests its oncogenic nature.

Taya et al. (80) have recently described a human lung giant cell carcinoma grown in nude mice, where both c-Ki-*ras* and c-*myc* oncogenes were amplified about 10-fold. Besides, sequencing studies indicated that at least some of the amplified c-Ki-*ras* copies were also mutationally activated in the 12th codon. These results fit to the multistage theory of cancer development and progression (see 58). Apparently co-operating lesions in cellular oncogenes accumulate during tumour growth and selection and increase the malignant potential of the tumour cells (44).

When does oncogene amplification come into play during tumourigenesis? Gene amplification may not be any initiating event in carcinogenesis. Amplification and enhanced expression of c-*myc* and N-*myc* may occur during the progression of human carcinoma of the lung and neuroblastoma cells to a more malignant phenotype (14, 53, 73). There may be, however, no mandatory sequence of oncogene amplifications for the genesis of any particular tumor. Amplification of an oncogene could play its part in malignant progression of already initiated cells whenever it happened to occur.

ACKNOWLEDGEMENTS

I thank my colleagues Manfred Schwab, Gerard Evan, J. Michael Bishop, Robert Winqvist, Kalle Saksela, Jorma Keski-Oja, C. C. Lin, Arthur Levinson, Wendy Colby and Donna George for collaboration, Ron Ellis for monoclonal antibodies against p21 proteins and Stephen Hann and Robert Eisenman for communicating their results before publication. The studies in the author's laboratory were supported by the Finnish Cancer Research Fund and by the Academy of Finland.

REFERENCES

1. *Abrams HD, Rohrschneider LR, Eisenman RN:* Nuclear location of the putative transforming protein of avian myelocytomatosis virus. Cell 29: 427–439, 1982
2. *Alitalo K, Bishop JM, Smith DH, Chen EY, Colby WW, Lenvinson AD:* Nucleotide sequence of the v-*myc* oncogene of avian retrovirus MC29. Proc Natl Acad Sci USA 80: 100–104, 1983
3. *Alitalo K, Ramsay G, Bishop JM, Ohlsson-Pfeifer S, Colby WW, Levinson AD:* Identification of nuclear proteins encoded by viral and cellular *myc* oncogenes. Nature 306: 274–277, 1983
4. *Alitalo K, Saksela K, Winqvist R, Schwab M, Bishop JM:* Amplification and aberrant expression of cellular oncogenes in human colon cancer cells. In: Genes and cancer. Ed. J. M. Bishop and J. Rowley. Alan Liss Co, New York, in press
5. *Alitalo K, Schwab M, Lin CC, Varmus HE, Bishop JM:* Homogeneously staining chromosomal regions con-

tain amplified copies of an abundantly expressed cellular oncogene (c-*myc*) in malignant neuroendocrine cells from a human colon carcinoma. Proc Natl Acad Sci USA 80: 1707—1711, 1983

6. *Alitalo K, Winqvist R, Lin CC, de la Chapelle A, Schwab M, Bishop JM:* Aberrant expression of an amplified c-*myb* oncogene in two cell lines from a colon carcinoma. Proc Natl Acad Sci USA 81: 4534—4538, 1984
7. *Armelin HA, Armelin MCS, Kelly K, Stewart T, Leder P, Cochran BH, Stiles CD:* Functional role for c-*myc* in mitogenic response to platelet-derived growth factor. Nature 310: 655—660, 1984
8. *Barker PE:* Double Minutes in human tumor cells. Cancer Genet Cytogenet 5: 81—94, 1982
9. *Barker PE, Drwinga HL, Mittelman WN, Maddox AM:* Double minutes replicate once during S phase of the cell cycle. Exp Cell Res 130: 353—360, 1980
10. *Barsoum J, Varshavsky A:* Mitogenic hormones and tumor promoters greatly increase the incidence of colony-forming cells bearing amplified dihydrofolate reductase genes. Proc Natl Acad Sci USA 80: 5330—5334, 1983
11. *Biedler JL, Meyers MB, Spengler BA:* Homogeneously staining regions and double minute chromosomes, prevalent cytogenetic abnormalities of human neuroblastoma cells. Adv Cell Neurobiol 4: 268—301, 1983
12. *Biedler JL, Spengler BA:* Metaphase chromosome anomaly: association with drug resistance and cell-specific products. Science 191: 185—187, 1976
13. *Brodeur GM, Green AA, Hayes FA, Williams KJ, Williams DL, Tsiatis AA:* Cytogenetic features of human neuroblastomas and cell lines. Cancer Res 41: 4678—4686, 1981
14. *Brodeur GM, Seeger RC, Schwab M, Varmus HE, Bishop JM:* Amplification of N-*myc* in Untreated Human Neuroblastomas Correlates with Advanced Disease Stage. Science 224: 1121—1124, 1984
15. *Brown PC, Tlsty TD, Schimke RT:* Enhancement of methotrexate resistance and DHFR gene amplification by treatment of 3T6 cells with hydroxyurea. Mol Cell Biol 3: 1097—1107, 1983
16. *Bullock P, Botchan M:* Molecular events in the excision of SV40 DNA from the chromosomes of cultured mammalian cells. In: Gene amplification. Ed. R. T. Schimke. Cold Spring Harbor Lab 215—230, 1982
17. *Calabretta B, Robberson DL, Barrera-Saldana HA, Lambrou TP, Saunders GF:* Genome instability in a region of human DNA enriched in Alu repeat sequences. Nature 296: 219—225, 1982
18. *Chattopadhyay SK, Chang EH, Lander MR, Ellis RW, Scolnick EM, Lowy DR:* Amplification and rearrangement of *onc* genes in mammalian species. Nature 296: 361—363, 1982
19. *Cifone MA, Fidler IJ:* Increasing metastatic potential is associated with increasing genetic instability of clones isolated from murine neoplasms. Proc Natl Acad Sci USA 78: 6949—6952, 1981
20. *Collins S, Groudine M:* Amplification of endogenous *myc*-related DNA sequences in a human myeloid leukemia cell line. Nature 298: 679—681, 1982
21. *Collins SJ, Groudine MT:* Rearrangement and amplification of c-*abl* sequences in the human chronic myelogenous leukemia cell line K-562. Proc Natl Acad Sci USA 80: 4813—4817, 1983
22. *Collins SJ, Kubonishi I, Miyoshi I, Groudine MT:* Altered transcription of the c-*abl* oncogene in K-562 and other chronic myelogenous leukemia cells. Science 225: 72—74, 1984
23. *Cowell JK:* Double minutes and homogeneously staining regions: Gene amplification in mammalian cells. Ann Rev Genet 16: 21—52, 1982
24. *Cox D, Yuncken C, Spriggs AI:* Numerous minute chromatin bodies in malignant tumours of childhood. Lancet I: 55—58, 1965
25. *Dalla-Favera RD, Wong-Staal F, Gallo RG: Onc* gene amplification in promyelocytic leukemia cell line HL-60 and primary leukaemic cells of the same patient. Nature 299: 61—63, 1982
26. *Donner P, Greiser-Wilke I, Moelling K:* Nuclear Localization and DNA binding of the transforming gene product of avian myelocytomatosis virus. Nature 296: 262—266, 1982
27. *Downward J, Yarden Y, Mayes E, Scrace G, Totty N, Stockwell P, Ullrich A, Schlessinger J, Waterfield MD:* Close similarity of epidermal growth factor receptor and v-*erb*B oncogene protein sequences. Nature 307: 521—527, 1984
28. *Echols H:* SOS-functions, cancer and inducible evolution. Cell 25: 1—2, 1981
29. *Eisenman RN, Tachibana CY, Abrams HD, Hann SR:* v-*myc* and c-*myc*-Encoded proteins are associated with the nuclear matrix. Submitted
30. *Gilbert F, Balaban G, Brangman D, Herrmann N, Lister A:* Homogeneously staining regions and tumorigenicity. Int J Cancer 31: 765—768, 1983
31. *Gill GN, Lazar CS:* Increased phosphotyrosine content and inhibition of proliferation in EGF-treated A431 cells. Nature 293: 305—307, 1981
32. *Hann SR, Abrams HD, Rohrschneider LR, Eisenman RN:* Proteins encoded by v-*myc* and c-*myc* oncogenes: Identification and localization in acute leukemia virus transformants and bursal lymphoma cell lines. Cell 34: 789—798, 1983
33. *Hann SR, Eisenman RN:* Two proteins encoded by the human c-*myc* oncogene: Differential expression in neoplastic cells. Mol Cell Biol 4: 2486—2497, 1984
34. *Harper ME, Franchini G, Love J, Simon MI, Gallo RC, Wong-Staal F:* Chromosomal sublocalization of human c-*myb* and c-*fes* cellular onc genes. Nature 304: 169—171, 1983
35. *Heisterkamp N, Stephenson JR, Groffen J, Hansen PF, de Klein A, Bartman CR, Grosveld G:* Localization of the c-*abl* oncogene adjacent to a translocation break point in chronic myelocytic leukaemia. Nature 306: 239—242, 1983
36. *Heldin C-H, Westermark B:* Growth factors: Mechanism of action and relation to oncogenes. Cell 37: 9—20, 1984
37. *Johnston RN, Beverley SM, Schimke RT:* Rapid spontaneous dihydrofolate reductase gene amplification shown by fluorescence-activated cell sorting. Proc Natl Acad Sci USA 80: 3711—3715, 1983
38. *Kanda N, Schreck R, Alt F, Bruns G, Baltimore D, Latt S:* Isolation of amplified DNA sequences from IMR-32 human neuroblastoma cells: Facilitation by fluoroscence-activated flow sorting of metaphase chromosomes. Proc Natl Acad Sci USA 80: 4069—4073, 1983
39. *Kelly K, Cochran BH, Stiles CD, Leder P:* Cell-specific regulation of the c-*myc* gene by lymphocyte mitogens and platelet-derived growth factor. Cell 35: 603—610, 1983
40. *Klein G:* Specific chromosome translocations and the genesis of B-cell derived tumors in mice and men. Cell 32: 311—315, 1983
41. *Konopka JB, Watanabe SM, Witte ON:* An alteration of the human c-*abl* protein in K562 leukemia cells unmasks associated tyrosine kinase activity. Cell 37: 1035—1042, 1984
42. *Kovacs G:* Homogeneously staining regions on marker

chromosomes in malignancy. Int J Cancer 23: 299–301, 1979
43. *Kozbor D, Croce CM:* Amplification of the c-*myc* oncogene in one of five human breast carcinoma cell lines. Cancer Res 44: 438–441, 1984
44. *Land H, Parada LF, Weinberg RA:* Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes. Science 222: 771–778, 1983
45. *Latt SA:* Sister chromatid exchange formation. Ann Rev Genet 15: 11–15, 1981
46. *Laughlin TJ, Taylor JH:* Initiation of DNA replication in chromosomes of Chinese hamster ovary cells. Chromosoma 75: 19–35, 1979
47. *Lavi S:* Carcinogen-mediated amplification of viral DNA sequences in simian virus 40-transformed Chinese hamster embryo cells. Proc Natl Acad Sci USA 78: 6144–6148, 1981
48. *Lee W-H, Murphree AL, Benedict WF:* Expression and amplification of the N-*myc* gene in primary retinoblastoma. Nature 309: 458–460, 1984
49. *Levan A, Manolov G, Clifford P:* Chromosomes of a human neuroblastoma. A new case with accessory minute chromosomes. J Natl Cancer Inst 41: 1377–1387, 1968
50. *Levan A, Levan G, Mitelman F:* Chromosomes and cancer. Hereditas 86: 15–29, 1977
51. *Lewis JA, Biedler JL, Melera PW:* Gene amplification accompanies low level increases in the activity of dihydrofolate reductase in antifolate-resistant Chinese hamster lung cells containing abnormally banding chromosomes. J Cell Biol 94: 418–424, 1982
52. *Lin CC, Alitalo K, Schwab M, George D, Varmus HE, Bishop JM:* Evolution of karyotypic abnormalities and c-*myc* oncogene amplification in a human colonic carcinoma. Submitted for publication
53. *Little CD, Nau MM, Carney DN, Gazdar AF, Minna JD:* Amplification an expression of the c-*myc* oncogene in human lung cancer cell lines. Nature 306: 194–196, 1983
54. *Lozzio CB, Lozzio BB:* Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. Blood 45: 321–334, 1975
55. *Mariani BD, Schimke RT:* Gene amplification in a single cell cycle in Chinese hamster ovary cells. J Biol Chem 259: 1901–1910, 1984
56. *Mark J:* Double minutes-a chromosomal aberration in Rous sarcomas in mice. Hereditas 57: 1–22, 1967
57. *Montgomery KT, Biedler JL, Spengler BA, Melera PW:* Specific DNA sequence amplification in human neuroblastoma cells. Proc Natl Acad Sci USA 80: 5724–5728, 1983
58. *Nowell RC:* The clonal evolution of tumor cell populations. Science 194: 23–28, 1976
59. *Nowell R, Finan J, Favera RD, Gallo RC, Ar-Rushdi A, Romanczuk G, Selden JR, Emanuel BS, Rovera G, Croce CM:* Association of amplified oncogene c-*myc* with an abnormally banded chromosome 8 in a human leukemia cell line. Nature 306: 494–497, 1983
60. *Pall ML:* Gene amplification model of carcinogenesis. Proc Natl Acad Sci USA 78: 2465–2468, 1981
61. *Pelicci P-G, Lanfrancone L, Brathwaite MD, Wolman SR, Dalla-Favera R:* Amplification of the c-*myb* oncogene in a case of human acute myelogenous leukemia. Science 224: 1117–1121, 1984
62. *Pellegrini S, Dailey L, Basilico C:* Amplification and excision of integrated polyoma DNA sequences require a functional origin of replication. Cell. 36: 943–949, 1984
63. *Quinn LA, Moore GE, Morgan RT, Woods LK:* Cell lines from human colon carcinoma with unusual cell products, double minutes, and homogeneously staining regions. Cancer Res 39: 4914, 1979
64. *Rechavi G, Givol D, Canaani E:* Activation of a cellular oncogene by DNA rearrangement: possible involvement of an IS-like element Nature 300: 607–611, 1982
65. *Roberts JM, Buck LB, Axel R:* A structure for amplified DNA. Cell 33: 53–63, 1983
66. *Rowley JD:* Human oncogene locations and chromosome aberrations. Nature 301: 290, 1983
67. *Rowley JD, Testa JR:* Chromosome abnormalities in malignant hematologic diseases. Adv Cancer Res 36: 103–147, 1982
68. *Rowley JD, Ultmann JE eds:* Chromosomes and Cancer: from molecules to man. Academic Press, New York 1983
69. *Saksela K, Bergh J, Lehto V-P, Nilsson K, Alitalo K:* Amplification of the c-*myc* oncogene is characteristic of a subpopulation of human small cell lung cancer. Cancer Res, in press
70. *Schimke RT:* Gene amplification. Cold Spring Harbor Labor 1982
71. *Schimke RT, Brown PC, Kaufman RJ, McGrogan M, Slate DL:* Chromosomal and extrachromosomal localization of amplified dihydrofolate reductase genes in cultured mammalian cells. Cold Spring Harbor Symp Quant Biol 55: 785–797, 1981
72. *Schwab M, Alitalo K, Klempnauer K-H, Varmus HE, Bishop JM, Gelbert F, Brodeur G, Goldstein M, Trent J:* Amplified DNA with limited homology to *myc* cellular oncogne is shared by human neuroblastoma cell lines and a neuroblastoma tumour. Nature 305: 245–248, 1983
73. *Schwab M, Alitalo K, Varmus HE, Bishop JM:* Amplification of cellular oncogenes in tumor cells. In: The cancer cell. Cold Spring Harbor Press 2: 215–220, 1984
74. *Schwab M, Alitalo K, Varmus HE, Bishop JM, George D:* A cellular oncogene (c-Ki-*ras*) is amplified, overexpressed, and located within karyotypic abnormalities in mouse adreocortical tumour cells. Nature 303: 497–501, 1983
75. *Schwab M, Varmus HE, Bishop JM, Grezeschik K-H, Naylor SL, Sakaguchi AY, Brodeur G, Trent J:* Chromosome localization in normal human cells and neuroblastomas of a gene related to c-*myc*. Nature 308: 288–291, 1984
76. *Selden JR, Emanuel BS, Wang E, Cannizzaro L, Palumbo A, Erikson J, Nowell PC, Rovera G, Croce CM:* Amplified C_λ and c-*abl* genes are on the same marker chromosome in K562 leukemia cells. Proc Natl Acad Sci USA 80: 7289–7292, 1983
77. *Shimizu K, Goldfarb M, Perucho M, Wigler M:* Isolation and preliminary characterization of the transforming gene of human neuroblastoma cell line. Proc Natl Acad Sci USA 80: 383–387, 1983
78. *Shmookler Reis RJ, Lumpkin CK, McGill JR, Riabowol TK, Goldstein S:* Extrachromosomal circular copies of an 'inter-Alu' unstable sequence in human DNA are amplified during in vitro and in vivo ageing. Nature 201: 394–398, 1983
79. *Stark GR, Wahl GM:* Gene amplification. Ann Rev Biochem 53: 447–491, 1984
80. *Taya Y, Hesogai K, Hirohashi S, Shimosato Y, Tsuchiya R, Tsuchida N, Fushimi M, Sekiya T, Nishimura S:* A novel combination of K-*ras* and *myc* amplification accompanied by point mutational activation of K-*ras* in human lung cancer. EMBO J 3: 2943–2946, 1984
81. *Tlsty TD, Brown PC, Schimke TR:* UV radiation facilitates methotrexate resistance and amplification of the dihydrofolate reductase gene in cultured 3T6 mouse cells. Molec Cell Biol 4: 1050–1056, 1984

82. *Ullrich A, Coussens L, Hayflick JS, Dull TJ, Gray A, Tam AW, Lee J, Yarden Y, Libermann TA, Sschlessinger J, Downward J, Mayes ELV, Whittle N, Waterfield MD, Seeburg PH:* Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells. Nature 309: 418—425, 1984
83. *Wahl GM, de Saint Vincent BR, DeRose ML:* Effect of chromosomal position on amplification of transfected genes in animal cells. Nature 307: 516—520, 1984
84. *Varshavsky A:* On the possibility of metabolic control of replicon 'misfiring': Relationship to emergence of malignant phenotypes in mammalian cell lineages. Proc Natl Acad Sci USA 78: 3673—3677, 1981
85. *Varshavsky A:* Phorbol ester dramatically increases incidence of methotrexate-resistant mouse cells: Possible mechanisms and relevance to tumor promotion. Cell 25: 561—572, 1981
86. *Whang-Peng J, Bunn PA, Kao-Shan CS, Lee EC, Carney DN, Gazdar A, Minna JD:* A Nonrandom chromosomal abnormality, del 3p (14—23), in human small cell lung cancer (SCLC). Cancer Genet Cytogenet 6: 119—134, 1982
87. *Whang-Peng J, Kao-Shan CS, Lee EC, Bunn PA, Carney DN, Gazdar AF, Portlock C, Minna JD:* Deletion 3p (14—23), Double minute chromosomes, and homogeneously staining regions in human small-cell lung cancer. sCSH Laboratory 1982
88. *Winqvist R, Knuutila S, Leprince D, Stehelin D, Alitalo K:* Mapping of amplified c-*myb* oncogene, sister chromatid exchanges and karyotypic analysis of the COLO 205 colon carcinoma cell line. Cancer Genet Cytogenet. in press
89. *Winqvist R, Saksela K, Alitalo K: myc* proteins are not associated with chromatin in mitotic cells. EMBO J 3: 2947—2950
90. *Woodcock DM, Cooper IA:* Evidence for double replication of chromosomal DNA segments as a general consequence of DNA replication inhibition. Cancer Res 41: 2483—2490, 1981
91. *Zabel BU, Naylor SL, Grezeschik K-H, Sakaguchi AY:* Regional assignment of human protooncogene c-*myb* to 6q21—qter. Somatic Cell Molec Genet 10: 105—108, 1984

Received for publication: September 19, 1984
Address: K. Alitalo
Department of Virology
University of Helsinki
SF-00290 Helsinki
Finland

EXHIBIT 3

MOLECULAR BIOLOGY OF THE CELL

THIRD EDITION

Text Editor: Miranda Robertson
Managing Editor: Ruth Adams
Illustrator: Nigel Orme
Molecular Model Drawings: Kate Hesketh-Moore
Director of Electronic Publishing: John M-Roblin
Computer Specialist: Chuck Bartelt
Disk Preparation: Carol Winter
Copy Editor: Shirley M. Cobert
Production Editor: Douglas Goertzen
Production Coordinator: Perry Bessas
Indexer: Maija Hinkle

Bruce Alberts received his Ph.D. from Harvard University and is currently President of the National Academy of Sciences and Professor of Biochemistry and Biophysics at the University of California, San Francisco. *Dennis Bray* received his Ph.D. from the Massachusetts Institute of Technology and is currently a Medical Research Council Fellow in the Department of Zoology, University of Cambridge. *Julian Lewis* received his D.Phil. from the University of Oxford and is currently a Senior Scientist in the Imperial Cancer Research Fund Developmental Biology Unit, University of Oxford. *Martin Raff* received his M.D. from McGill University and is currently a Professor in the MRC Laboratory for Molecular Cell Biology and the Biology Department, University College London. *Keith Roberts* received his Ph.D. from the University of Cambridge and is currently Head of the Department of Cell Biology, the John Innes Institute, Norwich. *James D. Watson* received his Ph.D. from Indiana University and is currently Director of the Cold Spring Harbor Laboratory. He is the author of *Molecular Biology of the Gene* and, with Francis Crick and Maurice Wilkins, won the Nobel Prize in Medicine and Physiology in 1962.

Library of Congress Cataloging-in-Publication Data
Molecular biology of the cell / Bruce Alberts . . . [et al.].—3rd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-8153-1619-4 (hard cover).—ISBN 0-8153-1620-8 (pbk.)
1. Cytology. 2. Molecular biology. I. Alberts, Bruce.
[DNLM: 1. Cells. 2. Molecular Biology. QH 581.2 M718 1994]
QH581.2.M64 1994
574.87—dc20
DNLM/DLC
for Library of Congress 93-45907
CIP

Published by Garland Publishing, Inc.
717 Fifth Avenue, New York, NY 10022

Printed in the United States of America

15 14 13 12 10 9 8 7

Front cover: The photograph shows a rat nerve cell in culture. It is labeled (*yellow*) with a fluorescent antibody that stains its cell body and dendritic processes. Nerve terminals (*green*) from other neurons (not visible), which have made synapses on the cell, are labeled with a different antibody. (Courtesy of Olaf Mundigl and Pietro de Camilli.)

Dedication page: Gavin Borden, late president of Garland Publishing, weathered in during his mid-1980s climb near Mount McKinley with MBoC author Bruce Alberts and famous mountaineer guide Mugs Stump (1940–1992).

Back cover: The authors, in alphabetical order, crossing Abbey Road in London on their way to lunch. Much of this third edition was written in a house just around the corner. (Photograph by Richard Olivier.)

extracts. If these minor cell proteins differ among cells to the same extent as the more abundant proteins, as is commonly assumed, only a small number of protein differences (perhaps several hundred) suffice to create very large differences in cell morphology and behavior.

A Cell Can Change the Expression of Its Genes in Response to External Signals [3]

Most of the specialized cells in a multicellular organism are capable of altering their patterns of gene expression in response to extracellular cues. If a liver cell is exposed to a glucocorticoid hormone, for example, the production of several specific proteins is dramatically increased. Glucocorticoids are released during periods of starvation or intense exercise and signal the liver to increase the production of glucose from amino acids and other small molecules; the set of proteins whose production is induced includes enzymes such as tyrosine aminotransferase, which helps to convert tyrosine to glucose. When the hormone is no longer present, the production of these proteins drops to its normal level.

Other cell types respond to glucocorticoids in different ways. In fat cells, for example, the production of tyrosine aminotransferase is reduced, while some other cell types do not respond to glucocorticoids at all. These examples illustrate a general feature of cell specialization—different cell types often respond in different ways to the same extracellular signal. Underlying this specialization are features that do not change, which give each cell type its permanently distinctive character. These features reflect the persistent expression of different sets of genes.

Gene Expression Can Be Regulated at Many of the Steps in the Pathway from DNA to RNA to Protein [4]

If differences between the various cell types of an organism depend on the particular genes that the cells express, at what level is the control of gene expression exercised? There are many steps in the pathway leading from DNA to protein, and all of them can in principle be regulated. Thus a cell can control the proteins it makes by (1) controlling when and how often a given gene is transcribed (**transcriptional control**), (2) controlling how the primary RNA transcript is spliced or otherwise processed (**RNA processing control**), (3) selecting which completed mRNAs in the cell nucleus are exported to the cytoplasm (**RNA transport control**), (4) selecting which mRNAs in the cytoplasm are translated by ribosomes (**translational control**), (5) selectively destabilizing certain mRNA molecules in the cytoplasm (**mRNA degradation control**), or (6) selectively activating, inactivating, or compartmentalizing specific protein molecules after they have been made (**protein activity control**) (Figure 9–2).

For most genes transcriptional controls are paramount. This makes sense because, of all the possible control points illustrated in Figure 9–2, only transcriptional control ensures that no superfluous intermediates are synthesized. In the

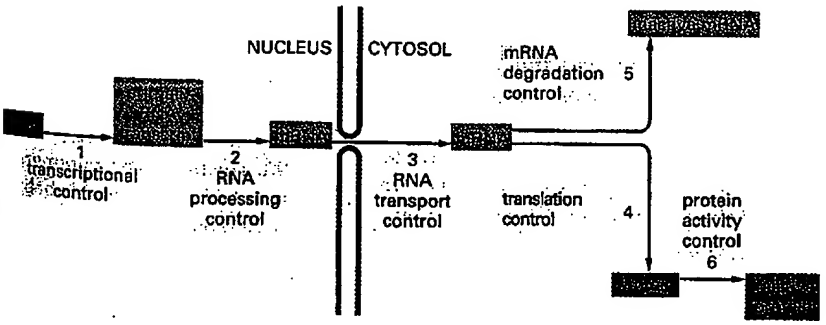

Figure 9–2 Six steps at which eucaryote gene expression can be controlled. Only controls that operate at steps 1 through 5 are discussed in this chapter. The regulation of protein activity (step 6) is discussed in Chapter 5; this includes reversible activation or inactivation by protein phosphorylation as well as irreversible inactivation by proteolytic degradation.

following sections we discuss the DNA and protein components that regulate the initiation of gene transcription. We return at the end of the chapter to the other ways of regulating gene expression.

Summary

The genome of a cell contains in its DNA sequence the information to make many thousands of different protein and RNA molecules. A cell typically expresses only a fraction of its genes, and the different types of cells in multicellular organisms arise because different sets of genes are expressed. Moreover, cells can change the pattern of genes they express in response to changes in their environment, such as signals from other cells. Although all of the steps involved in expressing a gene can in principle be regulated, for most genes the initiation of RNA transcription is the most important point of control.

DNA-binding Motifs in Gene Regulatory Proteins [5]

How does a cell determine which of its thousands of genes to transcribe? As discussed in Chapter 8, the transcription of each gene is controlled by a regulatory region of DNA near the site where transcription begins. Some regulatory regions are simple and act as switches that are thrown by a single signal. Other regulatory regions are complex and act as tiny microprocessors, responding to a variety of signals that they interpret and integrate to switch the neighboring gene on or off. Whether complex or simple, these switching devices consist of two fundamental types of components: (1) short stretches of DNA of defined sequence and (2) *gene regulatory proteins* that recognize and bind to them.

We begin our discussion of gene regulatory proteins by describing how these proteins were discovered.

Gene Regulatory Proteins Were Discovered Using Bacterial Genetics [6]

Genetic analyses in bacteria carried out in the 1950s provided the first evidence of the existence of **gene regulatory proteins** that turn specific sets of genes on or off. One of these regulators, the *lambda repressor,* is encoded by a bacterial virus, *bacteriophage lambda.* The repressor shuts off the viral genes that code for the protein components of new virus particles and thereby enables the viral genome to remain a silent passenger in the bacterial chromosome, multiplying with the bacterium when conditions are favorable for bacterial growth (see Figure 6–80). The lambda repressor was among the first gene regulatory proteins to be characterized, and it remains one of the best understood, as we discuss later. Other bacterial regulators respond to nutritional conditions by shutting off genes encoding specific sets of metabolic enzymes when they are not needed. The *lac repressor,* for example, the first of these bacterial proteins to be recognized, turns off the production of the proteins responsible for lactose metabolism when this sugar is absent from the medium.

The first step toward understanding gene regulation was the isolation of mutant strains of bacteria and bacteriophage lambda that were unable to shut off specific sets of genes. It was proposed at the time, and later proved, that most of these mutants were deficient in proteins acting as specific repressors for these sets of genes. Because these proteins, like most gene regulatory proteins, are present in small quantities, it was difficult and time-consuming to isolate them. They were eventually purified by fractionating cell extracts on a series of standard chromatography columns (see pp. 166–169). Once isolated, the proteins were shown to bind to specific DNA sequences close to the genes that they

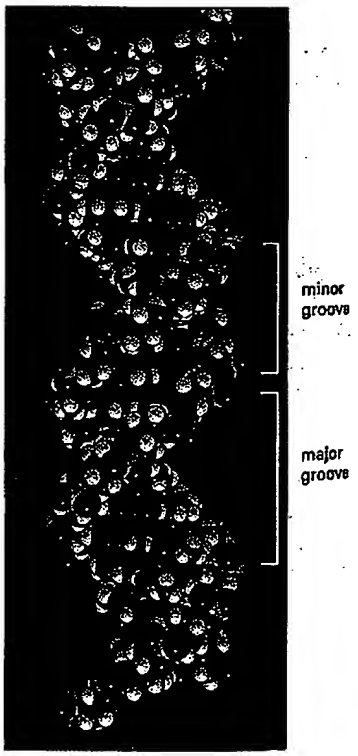

Figure 9–3 Double-helical structure of DNA. The major and minor grooves on the outside of the double helix are indicated. The atoms are colored as follows: carbon, *dark blue;* nitrogen, *light blue;* hydrogen, *white;* oxygen, *red;* phosphorus, *yellow.*

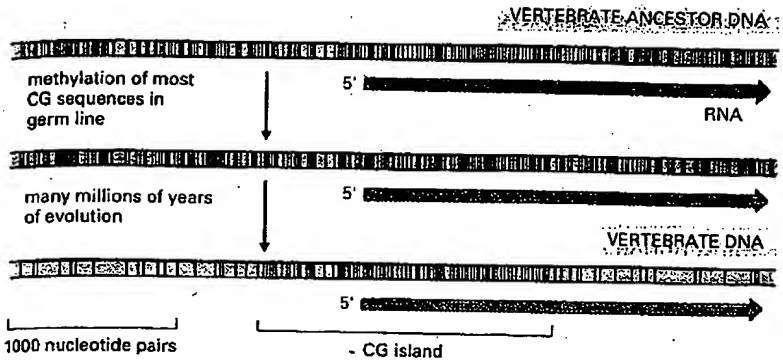

Figure 9–71 A mechanism to explain both the marked deficiency of CG sequences and the presence of CG islands in vertebrate genomes. A *black line* marks the location of an unmethylated CG dinucleotide in the DNA sequence, while a *red line* marks the location of a methylated CG dinucleotide.

Summary

The many types of cells in animals and plants are created largely through mechanisms that cause different genes to be transcribed in different cells. Since many specialized animal cells can maintain their unique character when grown in culture, the gene regulatory mechanisms involved in creating them must be stable once established and heritable when the cell divides, endowing the cell with a memory of its developmental history. Procaryotes and yeasts provide unusually accessible model systems in which to study gene regulatory mechanisms, some of which may be relevant to the creation of specialized cell types in higher eucaryotes. One such mechanism involves a competitive interaction between two (or more) gene regulatory proteins, each of which inhibits the synthesis of the other; this can create a flip-flop switch that switches a cell between two alternative patterns of gene expression. Direct or indirect positive feedback loops, which enable gene regulatory proteins to perpetuate their own synthesis, provide a general mechanism for cell memory.

In eucaryotes gene transcription is generally controlled by combinations of gene regulatory proteins. It is thought that each type of cell in a higher eucaryotic organism contains a specific combination of gene regulatory proteins that ensures the expression of only those genes appropriate to that type of cell. A given gene regulatory protein may be expressed in a variety of circumstances and typically is involved in the regulation of many genes.

In addition to diffusible gene regulatory proteins, inherited states of chromatin condensation are also utilized by eucaryotic cells to regulate gene expression. In vertebrates DNA methylation also plays a part, mainly as a device to reinforce decisions about gene expression that are made initially by other mechanisms.

Posttranscriptional Controls

Although controls on the initiation of gene transcription are the predominant form of regulation for most genes, other controls can act later in the pathway from RNA to protein to modulate the amount of gene product that is made. Although these **posttranscriptional controls,** which operate after RNA polymerase has bound to the gene's promoter and begun RNA synthesis, are less common than *transcriptional control,* for many genes they are crucial. It seems that every step in gene expression that could be controlled in principle is likely to be regulated under some circumstances for some genes.

We consider the varieties of posttranscriptional regulation in temporal order, according to the sequence of events that might be experienced by an RNA molecule after its transcription has begun (Figure 9–72).

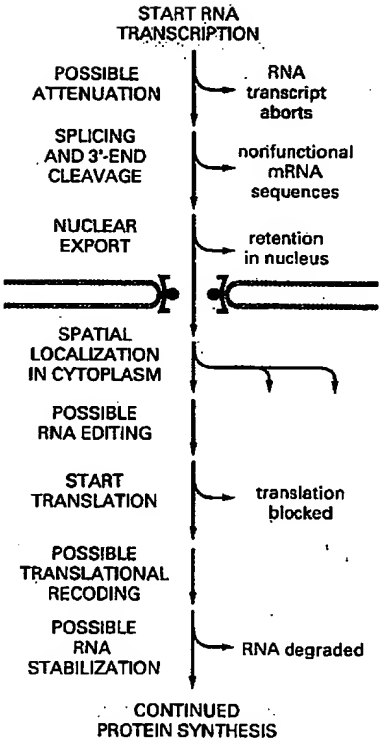

Figure 9–72 Possible posttranscriptional controls on gene expression. Only a few of these controls are likely to be used for any one gene.

EXHIBIT 4

MOLECULAR BIOLOGY OF THE CELL

fourth edition

Bruce Alberts

Alexander Johnson

Julian Lewis

Martin Raff

Keith Roberts

Peter Walter

GS Garland Science
Taylor & Francis Group

Garland
Vice President: Denise Schanck
Managing Editor: Sarah Gibbs
Senior Editorial Assistant: Kirsten Jenner
Managing Production Editor: Emma Hunt
Proofreader and Layout: Emma Hunt
Production Assistant: Angela Bennett
Text Editors: Marjorie Singer Anderson and Betsy Dilernia
Copy Editor: Bruce Goatly
Word Processors: Fran Dependahl, Misty Landers and Carol Winter
Designer: Blink Studio, London
Illustrator: Nigel Orme
Indexer: Janine Ross and Sherry Granum
Manufacturing: Nigel Eyre and Marion Morrow

Cell Biology Interactive
Artistic and Scientific Direction: Peter Walter
Narrated by: Julie Theriot
Production, Design, and Development: Mike Morales

Bruce Alberts received his Ph.D. from Harvard University and is President of the National Academy of Sciences and Professor of Biochemistry and Biophysics at the University of California, San Francisco. **Alexander Johnson** received his Ph.D. from Harvard University and is a Professor of Microbiology and Immunology at the University of California, San Francisco. **Julian Lewis** received his D.Phil. from the University of Oxford and is a Principal Scientist at the Imperial Cancer Research Fund, London. **Martin Raff** received his M.D. from McGill University and is at the Medical Research Council Laboratory for Molecular Cell Biology and Cell Biology Unit and in the Biology Department at University College London. **Keith Roberts** received his Ph.D. from the University of Cambridge and is Associate Research Director at the John Innes Centre, Norwich. **Peter Walter** received his Ph.D. from The Rockefeller University in New York and is Professor and Chairman of the Department of Biochemistry and Biophysics at the University of California, San Francisco, and an Investigator of the Howard Hughes Medical Institute.

Library of Congress Cataloging-in-Publicaton Data
Molecular biology of the cell / Bruce Alberts ... [et al.].-- 4th ed.
p. cm
Includes bibliographical references and index.
ISBN 0-8153-3218-1 (hardbound) -- ISBN 0-8153-4072-9 (pbk.)
1. Cytology. 2. Molecular biology. I. Alberts, Bruce.
[DNLM: 1. Cells. 2. Molecular Biology.]
QH581.2 .M64 2002
571.6--dc21

2001054471 CIP

Published by Garland Science, a member of the Taylor & Francis Group, 29 West 35th Street, New York, NY 10001-2299

Printed in the United States of America

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Front cover Human Genome: Reprinted by permission from *Nature,* International Human Genome Sequencing Consortium, 409:860–921, 2001 © Macmillan Magazines Ltd. Adapted from an image by Francis Collins, NHGRI; Jim Kent, UCSC; Ewan Birney, EBI; and Darryl Leja, NHGRI; showing a portion of Chromosome 1 from the initial sequencing of the human genome.

Back cover In 1967, the British artist Peter Blake created a design classic. Nearly 35 years later Nigel Orme (illustrator), Richard Denyer (photographer), and the authors have together produced an affectionate tribute to Mr Blake's image. With its gallery of icons and influences, its assembly created almost as much complexity, intrigue and mystery as the original. *Drosophila, Arabidopsis,* Dolly and the assembled company tempt you to dip inside where, as in the original, "a splendid time is guaranteed for all." (Gunter Blobel, courtesy of The Rockefeller University; Marie Curie, Keystone Press Agency Inc; Darwin bust, by permission of the President and Council of the Royal Society; Rosalind Franklin, courtesy of Cold Spring Harbor Laboratory Archives; Dorothy Hodgkin, © The Nobel Foundation, 1964; James Joyce, etching by Peter Blake; Robert Johnson, photo booth self-portrait early 1930s, © 1986 Delta Haze Corporation all rights reserved, used by permission; Albert L. Lehninger, (unidentified photographer) courtesy of The Alan Mason Chesney Medical Archives of The Johns Hopkins Medical Institutions; Linus Pauling, from Ava Helen and Linus Pauling Papers, Special Collections, Oregon State University; Nicholas Poussin, courtesy of ArtToday.com; Barbara McClintock, © David Micklos, 1983; Andrei Sakharov, courtesy of Elena Bonner; Frederick Sanger, © The Nobel Foundation, 1958.)

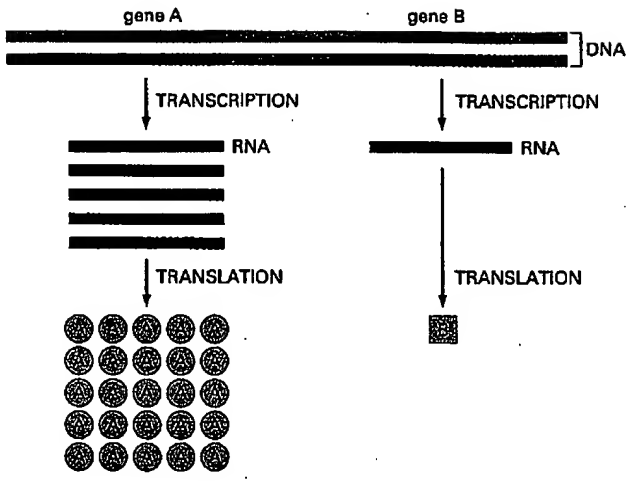

Figure 6–3 Genes can be expressed with different efficiencies. Gene A is transcribed and translated much more efficiently than gene B. This allows the amount of protein A in the cell to be much greater than that of protein B.

FROM DNA TO RNA

Transcription and translation are the means by which cells read out, or express, the genetic instructions in their genes. Because many identical RNA copies can be made from the same gene, and each RNA molecule can direct the synthesis of many identical protein molecules, cells can synthesize a large amount of protein rapidly when necessary. But each gene can also be transcribed and translated with a different efficiency, allowing the cell to make vast quantities of some proteins and tiny quantities of others (Figure 6–3). Moreover, as we see in the next chapter, a cell can change (or regulate) the expression of each of its genes according to the needs of the moment—most obviously by controlling the production of its RNA.

Portions of DNA Sequence Are Transcribed into RNA

The first step a cell takes in reading out a needed part of its genetic instructions is to copy a particular portion of its DNA nucleotide sequence—a gene—into an RNA nucleotide sequence. The information in RNA, although copied into another chemical form, is still written in essentially the same language as it is in DNA—the language of a nucleotide sequence. Hence the name **transcription.**

Like DNA, RNA is a linear polymer made of four different types of nucleotide subunits linked together by phosphodiester bonds (Figure 6–4). It differs from DNA chemically in two respects: (1) the nucleotides in RNA are *ribonucleotides*—that is, they contain the sugar ribose (hence the name *ribo*nucleic acid) rather than deoxyribose; (2) although, like DNA, RNA contains the bases adenine (A), guanine (G), and cytosine (C), it contains the base uracil (U) instead of the thymine (T) in DNA. Since U, like T, can base-pair by hydrogen-bonding with A (Figure 6–5), the complementary base-pairing properties described for DNA in Chapters 4 and 5 apply also to RNA (in RNA, G pairs with C, and A pairs with U). It is not uncommon, however, to find other types of base pairs in RNA: for example, G pairing with U occasionally.

Despite these small chemical differences, DNA and RNA differ quite dramatically in overall structure. Whereas DNA always occurs in cells as a double-stranded helix, RNA is single-stranded. RNA chains therefore fold up into a variety of shapes, just as a polypeptide chain folds up to form the final shape of a protein (Figure 6–6). As we see later in this chapter, the ability to fold into complex three-dimensional shapes allows some RNA molecules to have structural and catalytic functions.

Transcription Produces RNA Complementary to One Strand of DNA

All of the RNA in a cell is made by DNA transcription, a process that has certain similarities to the process of DNA replication discussed in Chapter 5.

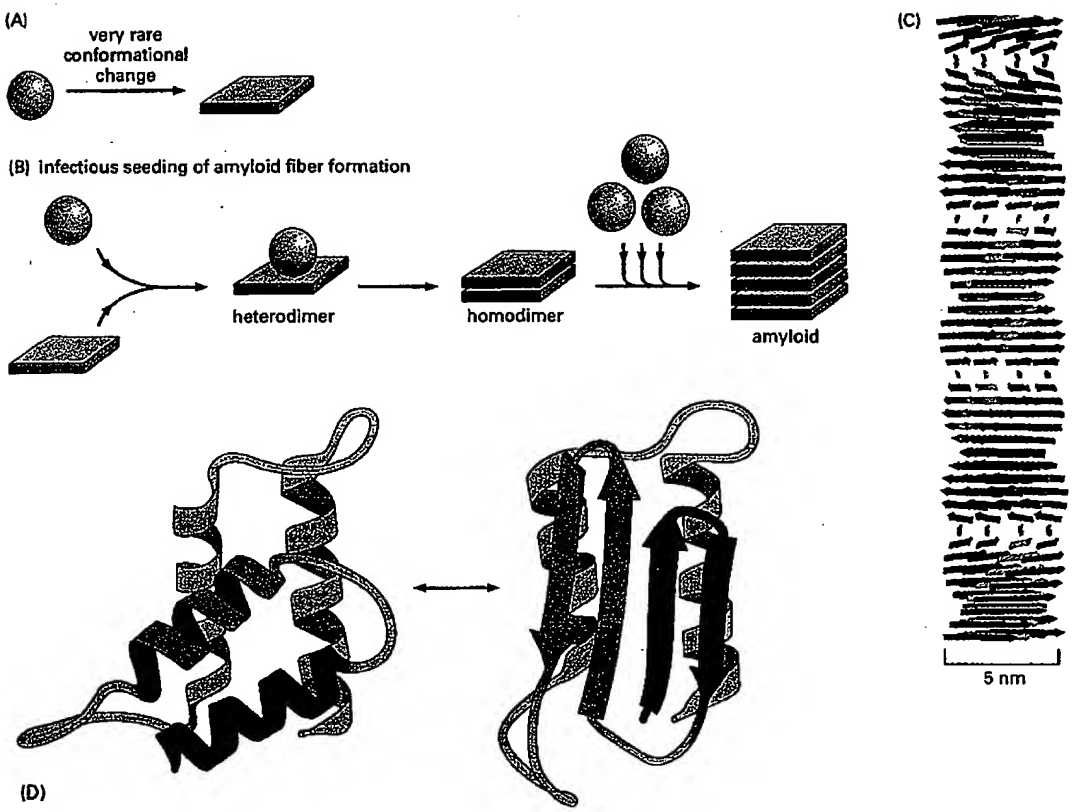

Figure 6–89 Protein aggregates that cause human disease. (A) Schematic illustration of the type of conformational change in a protein that produces material for a cross-beta filament. (B) Diagram illustrating the self-infectious nature of the protein aggregation that is central to prion diseases. PrP is highly unusual because the misfolded version of the protein, called PrP*, induces the normal PrP protein it contacts to change its conformation, as shown. Most of the human diseases caused by protein aggregation are caused by the overproduction of a variant protein that is especially prone to aggregation, but because this structure is not infectious in this way, it cannot spread from one animal to another. (C) Drawing of a cross-beta filament, a common type of protease-resistant protein aggregate found in a variety of human neurological diseases. Because the hydrogen-bond interactions in a β sheet form between polypeptide backbone atoms (see Figure 3–9), a number of different abnormally folded proteins can produce this structure. (D) One of several possible models for the conversion of PrP to PrP*, showing the likely change of two α-helices into four β-strands. Although the structure of the normal protein has been determined accurately, the structure of the infectious form is not yet known with certainty because the aggregation has prevented the use of standard structural techniques. (C, courtesy of Louise Serpell, adapted from M. Sunde et al., *J. Mol. Biol.* 273:729–739, 1997; D, adapted from S.B. Prusiner, *Trends Biochem. Sci.* 21:482–487, 1996.)

animals and humans. It can be dangerous to eat the tissues of animals that contain PrP*, as witnessed most recently by the spread of BSE (commonly referred to as the "mad cow disease") from cattle to humans in Great Britain.

Fortunately, in the absence of PrP*, PrP is extraordinarily difficult to convert to its abnormal form. Although very few proteins have the potential to misfold into an infectious conformation, a similar transformation has been discovered to be the cause of an otherwise mysterious "protein-only inheritance" observed in yeast cells.

There Are Many Steps From DNA to Protein

We have seen so far in this chapter that many different types of chemical reactions are required to produce a properly folded protein from the information contained in a gene (Figure 6–90). The final level of a properly folded protein in a cell therefore depends upon the efficiency with which each of the many steps is performed.

We discuss in Chapter 7 that cells have the ability to change the levels of their proteins according to their needs. In principle, any or all of the steps in Fig-

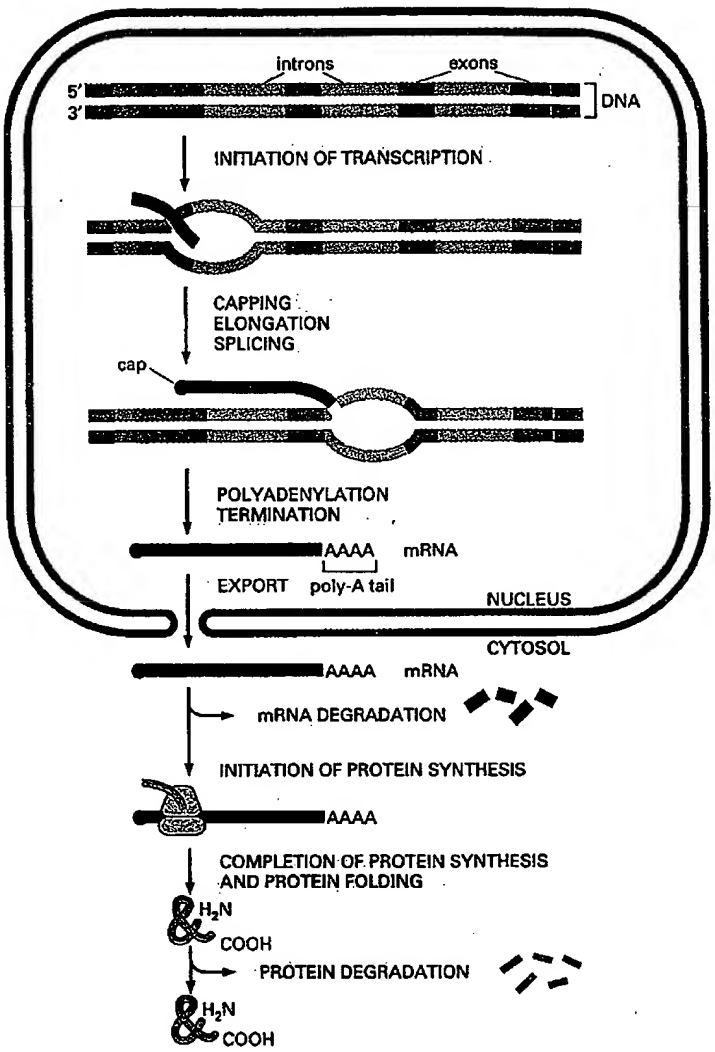

Figure 6–90 The production of a protein by a eucaryotic cell. The final level of each protein in a eucaryotic cell depends upon the efficiency of each step depicted.

ure 6–90) could be regulated by the cell for each individual protein. However, as we shall see in Chapter 7, the initiation of transcription is the most common point for a cell to regulate the expression of each of its genes. This makes sense, inasmuch as the most efficient way to keep a gene from being expressed is to block the very first step—the transcription of its DNA sequence into an RNA molecule.

Summary

The translation of the nucleotide sequence of an mRNA molecule into protein takes place in the cytoplasm on a large ribonucleoprotein assembly called a ribosome. The amino acids used for protein synthesis are first attached to a family of tRNA molecules, each of which recognizes, by complementary base-pair interactions, particular sets of three nucleotides in the mRNA (codons). The sequence of nucleotides in the mRNA is then read from one end to the other in sets of three according to the genetic code.

To initiate translation, a small ribosomal subunit binds to the mRNA molecule at a start codon (AUG) that is recognized by a unique initiator tRNA molecule. A large ribosomal subunit binds to complete the ribosome and begin the elongation phase of protein synthesis. During this phase, aminoacyl tRNAs—each bearing a specific amino acid bind sequentially to the appropriate codon in mRNA by forming complementary base pairs with the tRNA anticodon. Each amino acid is added to the C-terminal end of the growing polypeptide by means of a cycle of three sequential

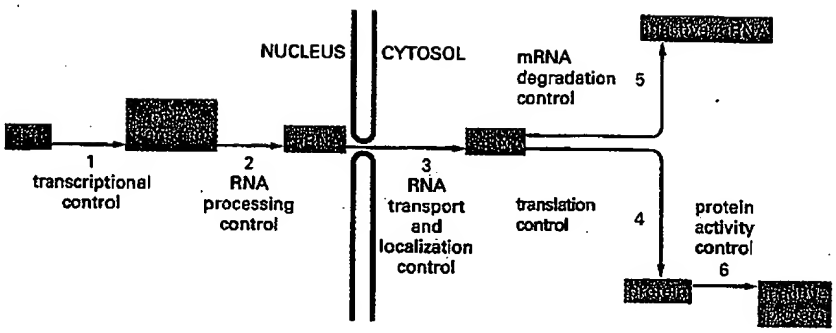

Figure 7–5 Six steps at which eucaryotic gene expression can be controlled. Controls that operate at steps 1 through 5 are discussed in this chapter. Step 6, the regulation of protein activity, includes reversible activation or inactivation by protein phosphorylation (discussed in Chapter 3) as well as irreversible inactivation by proteolytic degradation (discussed in Chapter 6).

Gene Expression Can Be Regulated at Many of the Steps in the Pathway from DNA to RNA to Protein

If differences among the various cell types of an organism depend on the particular genes that the cells express, at what level is the control of gene expression exercised? As we saw in the last chapter, there are many steps in the pathway leading from DNA to protein, and all of them can in principle be regulated. Thus a cell can control the proteins it makes by (1) controlling when and how often a given gene is transcribed (**transcriptional control**), (2) controlling how the RNA transcript is spliced or otherwise processed (**RNA processing control**), (3) selecting which completed mRNAs in the cell nucleus are exported to the cytosol and determining where in the cytosol they are localized (**RNA transport and localization control**), (4) selecting which mRNAs in the cytoplasm are translated by ribosomes (**translational control**), (5) selectively destabilizing certain mRNA molecules in the cytoplasm (**mRNA degradation control**), or (6) selectively activating, inactivating, degrading, or compartmentalizing specific protein molecules after they have been made (**protein activity control**) (Figure 7–5).

For most genes transcriptional controls are paramount. This makes sense because, of all the possible control points illustrated in Figure 7–5, only transcriptional control ensures that the cell will not synthesize superfluous intermediates. In the following sections we discuss the DNA and protein components that perform this function by regulating the initiation of gene transcription. We shall return at the end of the chapter to the additional ways of regulating gene expression.

Summary

The genome of a cell contains in its DNA sequence the information to make many thousands of different protein and RNA molecules. A cell typically expresses only a fraction of its genes, and the different types of cells in multicellular organisms arise because different sets of genes are expressed. Moreover, cells can change the pattern of genes they express in response to changes in their environment, such as signals from other cells. Although all of the steps involved in expressing a gene can in principle be regulated, for most genes the initiation of RNA transcription is the most important point of control.

DNA-BINDING MOTIFS IN GENE REGULATORY PROTEINS

How does a cell determine which of its thousands of genes to transcribe? As mentioned briefly in Chapters 4 and 6, the transcription of each gene is controlled by a regulatory region of DNA relatively near the site where transcription begins. Some regulatory regions are simple and act as switches that are thrown by a single signal. Many others are complex and act as tiny microprocessors, responding to a variety of signals that they interpret and integrate to switch the neighboring gene on or off. Whether complex or simple, these switching devices

occur in the germ line, the cell lineage that gives rise to sperm or eggs. Most of the DNA in vertebrate germ cells is inactive and highly methylated. Over long periods of evolutionary time, the methylated CG sequences in these inactive regions have presumably been lost through spontaneous deamination events that were not properly repaired. However promoters of genes that remain active in the germ cell lineages (including most housekeeping genes) are kept unmethylated, and therefore spontaneous deaminations of Cs that occur within them can be accurately repaired. Such regions are preserved in modern day vertebrate cells as CG islands. In addition, any mutation of a CG sequence in the genome that destroyed the function or regulation of a gene in the adult would be selected against, and some CG islands are simply the result of a higher than normal density of critical CG sequences.

The mammalian genome contains an estimated 20,000 CG islands. Most of the islands mark the 5′ ends of transcription units and thus, presumably, of genes. The presence of CG islands often provides a convenient way of identifying genes in the DNA sequences of vertebrate genomes.

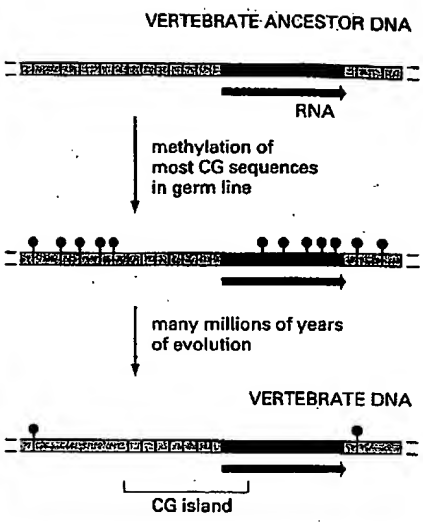

Figure 7–86 A mechanism to explain both the marked overall deficiency of CG sequences and their clustering into CG islands in vertebrate genomes. A *black line* marks the location of a CG dinucleotide in the DNA sequence, while a *red* "lollipop" indicates the presence of a methyl group on the CG dinucleotide. CG sequences that lie in regulatory sequences of genes that are transcribed in germ cells are unmethylated and therefore tend to be retained in evolution. Methylated CG sequences, on the other hand, tend to be lost through deamination of 5-methyl C to T, unless the CG sequence is critical for survival.

Summary

The many types of cells in animals and plants are created largely through mechanisms that cause different genes to be transcribed in different cells. Since many specialized animal cells can maintain their unique character through many cell division cycles and even when grown in culture, the gene regulatory mechanisms involved in creating them must be stable once established and heritable when the cell divides. These features endow the cell with a memory of its developmental history. Bacteria and yeasts provide unusually accessible model systems in which to study gene regulatory mechanisms. One such mechanism involves a competitive interaction between two gene regulatory proteins, each of which inhibits the synthesis of the other; this can create a flip-flop switch that switches a cell between two alternative patterns of gene expression. Direct or indirect positive feedback loops, which enable gene regulatory proteins to perpetuate their own synthesis, provide a general mechanism for cell memory. Negative feedback loops with programmed delays form the basis for cellular clocks.

In eucaryotes the transcription of a gene is generally controlled by combinations of gene regulatory proteins. It is thought that each type of cell in a higher eucaryotic organism contains a specific combination of gene regulatory proteins that ensures the expression of only those genes appropriate to that type of cell. A given gene regulatory protein may be active in a variety of circumstances and typically is involved in the regulation of many genes.

In addition to diffusible gene regulatory proteins, inherited states of chromatin condensation are also used by eucaryotic cells to regulate gene expression. An especially dramatic case is the inactivation of an entire X chromosome in female mammals. In vertebrates DNA methylation also functions in gene regulation, being used mainly as a device to reinforce decisions about gene expression that are made initially by other mechanisms. DNA methylation also underlies the phenomenon of genomic imprinting in mammals, in which the expression of a gene depends on whether it was inherited from the mother or the father.

POSTTRANSCRIPTIONAL CONTROLS

In principle, every step required for the process of gene expression could be controlled. Indeed, one can find examples of each type of regulation, although any one gene is likely to use only a few of them. Controls on the initiation of gene transcription are the predominant form of regulation for most genes. But other controls can act later in the pathway from DNA to protein to modulate the amount of gene product that is made. Although these **posttranscriptional controls**, which operate after RNA polymerase has bound to the gene's promoter and begun RNA synthesis, are less common than *transcriptional control*, for many genes they are crucial.

EXHIBIT 5

CHAPTER 29

Regulation of transcription

Genes VII (1997) CH29, pp. 847-848.
Benjamin Lewin

The phenotypic differences that distinguish the various kinds of cells in a higher eukaryote are largely due to differences in the expression of genes that code for proteins, that is, those transcribed by RNA polymerase II. In principle, the expression of these genes might be regulated at any one of several stages. The concept of the "level of control" implies that gene expression is not necessarily an automatic process once it has begun. It could be regulated in a gene-specific way at any one of several sequential steps. We can distinguish (at least) five potential control points, forming the series:

Activation of gene structure
↓
Initiation of transcription
↓
Processing the transcript
↓
Transport to cytoplasm
↓
Translation of mRNA

The existence of the first step is implied by the discovery that genes may exist in either of two structural conditions. Relative to the state of most of the genome, genes are found in an "active" state in the cells in which they are expressed (see Chapter 27). The change of structure is distinct from the act of transcription, and indicates that the gene is "transcribable." This suggests that acquisition of the "active" structure must be the first step in gene expression.

Transcription of a gene in the active state is controlled at the stage of initiation, that is, by the interaction of RNA polymerase with its promoter. This is now becoming susceptible to analysis in the *in vitro* systems (see Chapter 28). For most genes, this is a major control point: probably it is the most common level of regulation.

There is at present no evidence for control at subsequent stages of transcription in eukaryotic cells, for example, via antitermination mechanisms.

The primary transcript is modified by capping at the 5′ end, and usually also by polyadenylation at the 3′ end. Introns must be spliced out from the transcripts of interrupted genes. The mature RNA must be exported from the nucleus to the cytoplasm. Regulation of gene expression by selection of sequences at the level of nuclear RNA might involve any or all of these stages, but the one for which we have most evidence concerns changes in splicing: some genes are expressed by means of alternative splicing patterns whose regulation controls the type of protein product (see Chapter 30).

Finally, the translation of an mRNA in the cytoplasm can be specifically controlled. There is little evidence for the employment of this mechanism in adult somatic cells, but it does occur in some embryonic situations, as described in Chapter 7. The mechanism is presumed to involve the blocking of initiation of translation of some mRNAs by specific protein factors.

But having acknowledged that control of gene expression can occur at multiple stages, and that production of RNA cannot inevitably be equated with production of protein; it is clear

that the overwhelming majority of regulatory events occur at the initiation of transcription. Regulation of tissue-specific gene transcription lies at the heart of eukaryotic differentiation; indeed, we see examples in Chapter 38 in which proteins that regulate embryonic development prove to be transcription factors. A regulatory transcription factor serves to provide common control of a large number of target genes, and we seek to answer two questions about this mode of regulation: what identifies the common target genes to the transcription factor; and how is the activity of the transcription factor itself regulated in response to intrinsic or extrinsic signals?

Response elements identify genes under common regulation

The principle that emerges from characterizing groups of genes under common control is that *they share a promoter element that is recognized by a regulatory transcription factor.* An element that causes a gene to respond to such a factor is called a **response element**; examples are the **HSE** (heat shock response element), **GRE** (glucocorticoid response element), **SRE** (serum response element).

The properties of some inducible transcription factors and the elements that they recognize are summarized in **Table 29.1**. Response elements have the same general characteristics as upstream elements of promoters or enhancers. They contain short consensus sequences, and copies of the response elements found in different genes are closely related, but not necessarily identical. The region bound by the factor extends for a short distance on either side of the consensus sequence. In promoters, the elements are not present at fixed distances from the startpoint, but are usually <200 bp upstream of it. The presence of a single element usually is sufficient to confer the regulatory response, but sometimes there are multiple copies.

Response elements may be located in promoters or in enhancers. Some types of elements are typically found in one rather than the other: usually an HSE is found in a promoter, while a GRE is found in an enhancer. We assume that all response elements function by the same general principle. *A gene is regulated by a sequence at the promoter or enhancer that is recognized by a specific protein. The protein functions as a transcription factor needed for RNA polymerase to initiate. Active protein is available only under conditions when the gene is to be expressed; its absence means that the promoter is not activated by this particular circuit.*

An example of a situation in which many genes are controlled by a single factor is provided by the heat shock response. This is common to a wide range of prokaryotes and eukaryotes and involves multiple controls of gene expression: an increase in temperature turns off transcription of some genes, turns on transcription of the heat shock genes, and causes changes in the translation of mRNAs. The control of the heat shock genes illustrates the differences between prokaryotic and eukaryotic modes of control. In bacteria, a new sigma factor is synthesized that directs RNA polymerase holoenzyme to recognize an alter-

Table 29.1 Inducible transcription factors bind to response elements that identify groups of promoters or enhancers subject to coordinate control.

Regulatory Agent	Module	Consensus	Factor
Heat shock	HSE	CNNGAANNTCCNNG	HSTF
Glucocorticoid	GRE	TGGTACAAATGTTCT	Receptor
Phorbol ester	TRE	TGACTCA	AP1
Serum	SRE	CCATATTAGG	SRF

World Journal of Surgical Oncology

BioMed Central

Research

Open Access

Prostate stem cell antigen (PSCA) expression in human prostate cancer tissues and its potential role in prostate carcinogenesis and progression of prostate cancer

Zhao Zhigang*[1] and Shen Wenlv[2]

Address: [1]Department of Urology, Shantou University Medical College, Shantou, Guangdong, China and [2]Department of Urology, No 2. Affiliated Hospital of Shantou University Medical College, Shantou, Guangdong, China

Email: Zhao Zhigang* - zgzhaodr@yahoo.com.cn; Shen Wenlv - wlshen99@hotmail.com

* Corresponding author

Published: 10 May 2004

World Journal of Surgical Oncology 2004, 2:13

Received: 30 March 2004

Accepted: 10 May 2004

This article is available from: http://www.wjso.com/content/2/1/13

Abstract

Background: Prostate stem cell antigen (PSCA) is a recently defined homologue of the Thy-1/Ly-6 family of glycosylphosphatidylinositol (GPI)-anchored cell surface antigens. The purpose of the present study was to examine the expression status of PSCA protein and mRNA in clinical specimens of human prostate cancer (Pca) and to validate it as a potential molecular target for diagnosis and treatment of Pca.

Materials and Methods: Immunohistochemical (IHC) and *in situ* hybridization (ISH) analyses of PSCA expression were simultaneously performed on paraffin-embedded sections from 20 benign prostatic hyperplasia (BPH), 20 prostatic intraepithelial neoplasm (PIN) and 48 prostate cancer (Pca) tissues, including 9 androgen-independent prostate cancers. The level of PSCA expression was semiquantitatively scored by assessing both the percentage and intensity of PSCA-positive staining cells in the specimens. Then compared PSCA expression between BPH, PIN and Pca tissues and analysed the correlations of PSCA expression level with pathological grade, clinical stage and progression to androgen-independence in Pca.

Results: In BPH and low grade PIN, PSCA protein and mRNA staining were weak or negative and less intense and uniform than that seen in HGPIN and Pca. There were moderate to strong PSCA protein and mRNA expression in 8 of 11 (72.7%) HGPIN and in 40 of 48 (83.4%) Pca specimens examined by IHC and ISH analyses, with statistical significance compared with BPH (20%) and low grade PIN (22.2%) samples ($p < 0.05$, respectively). The expression level of PSCA increased with high Gleason grade, advanced stage and progression to androgen-independence ($p < 0.05$, respectively). In addition, IHC and ISH staining showed a high degree of correlation between PSCA protein and mRNA overexpression.

Conclusions: Our data demonstrate that PSCA as a new cell surface marker is overexpressed by a majority of human Pca. PSCA expression correlates positively with adverse tumor characteristics, such as increasing pathological grade (poor cell differentiation), worsening clinical stage and androgen-independence, and speculatively with prostate carcinogenesis. PSCA protein overexpression results from upregulated transcription of PSCA mRNA. PSCA may have prognostic utility and may be a promising molecular target for diagnosis and treatment of Pca.

Introduction

Prostate cancer (Pca) is the second leading cause of cancer-related death in American men and is becoming a common cancer increasing in China. Despite recently great progress in the diagnosis and management of localized disease, there continues to be a need for new diagnostic markers that can accurately discriminate between indolent and aggressive variants of Pca. There also continues to be a need for the identification and characterization of potential new therapeutic targets on Pca cells. Current diagnostic and therapeutic modalities for recurrent and metastatic Pca have been limited by a lack of specific target antigens of Pca.

Although a number of prostate-specific genes have been identified (i.e. prostate specific antigen, prostatic acid phosphatase, glandular kallikrein 2), the majority of these are secreted proteins not ideally suited for many immunological strategies. So, the identification of new cell surface antigens is critical to the development of new diagnostic and therapeutic approaches to the management of Pca.

Reiter RE et al [1] reported the identification of prostate stem cell antigen (PSCA), a cell surface antigen that is predominantly prostate specific. The PSCA gene encodes a 123 amino acid glycoprotein, with 30% homology to stem cell antigen 2 (Sca 2). Like Sca-2, PSCA also belongs to a member of the Thy-1/Ly-6 family and is anchored by a glycosylphosphatidylinositol (GPI) linkage. mRNA *in situ* hybridization (ISH) localized PSCA expression in normal prostate to the basal cell epithelium, the putative stem cell compartment of prostatic epithelium, suggesting that PSCA may be a marker of prostate stem/progenitor cells.

In order to examine the status of PSCA protein and mRNA expression in human Pca and validate it as a potential diagnostic and therapeutic target for Pca, we used immunohistochemistry (IHC) and *in situ* hybridization (ISH) simultaneously, and conducted PSCA protein and mRNA expression analyses in paraffin-embedded tissue specimens of benign prostatic hyperplasia (BPH, n = 20), prostate intraepithelial neoplasm (PIN, n = 20) and prostate cancer (Pca, n = 48). Furthermore, we evaluated the possible correlation of PSCA expression level with Pca tumorigenesis, grade, stage and progression to androgen-independence.

Materials and methods

Tissue samples

All of the clinical tissue specimens studied herein were obtained from 80 patients of 57–84 years old by prostatectomy, transurethral resection of prostate (TURP) or biopsies. The patients were classified as 20 cases of BPH, 20 cases of PIN, 40 cases of primary Pca, including 9 patients with recurrent Pca and a history of androgen ablation therapy (orchiectomy and/or hormonal therapy), who were referred to as androgen-independent prostate cancers. Eight specimens were harvested from these androgen-independent Pca patients prior to androgen ablation treatment. Each tissue sample was cut into two parts, one was fixed in 10% formalin for IHC and the other treated with 4% paraformaldehyde/0.1 M PBS PH 7.4 in 0.1% DEPC for 1 h for ISH analysis, and then embedded in paraffin. All paraffin blocks examined were then cut into 5 μm sections and mounted on the glass slides specific for IHC and ISH respectively in the usual fashion. H&E-stained section of each Pca was evaluated and assigned a Gleason score by the experienced urological pathologist at our institution based on the criteria of Gleason score [2]. The Gleason sums are summarized in Table 1. Clinical staging was performed according to Jewett-whitmore-prout staging system, as shown in Table 2. In the category of PIN, we graded the specimens into two groups, i.e. low grade PIN (grade I - II) and high grade PIN (HGPIN, grade III) on the basis of literatures [3,4].

Immunohistochemical (IHC) analysis

Briefly, tissue sections were deparaffinized, dehydrated, and subjected to microwaving in 10 mmol/L citrate buffer, PH 6.0 (Boshide, Wuhan, China) in a 900 W oven for 5 min to induce epitope retrieval. Slides were allowed to cool at room temperature for 30 min. A primary mouse antibody specific to human PSCA (Boshide, Wuhan, China) with a 1:100 dilution was applied to incubate with the slides at room temperature for 2 h. Labeling was detected by sequentially adding biotinylated secondary antibodies and strepavidin-peroxidase, and localized using 3,3'-diaminobenzidine reaction. Sections were then counterstained with hematoxylin. Substitution of the primary antibody with phosphate-buffered-saline (PBS) served as a negative-staining control.

mRNA in situ hybridization (ISH)

Five-μm-thick tissue sections were deparaffinized and dehydrated, then digested in pepsin solution (4 mg/ml in 3% citric acid) for 20 min at 37.5°C, and further processed for ISH. Digoxigenin-labeled sense and antisense human PSCA RNA probes (obtained from Boshide, Wuhan, China) were hybridized to the sections at 48°C overnight. The posthybridization wash with a high stringency was performed sequentially at 37°C in 2 × standard saline citrate (SSC) for 10 min, in 0.5 × SSC for 15 min and in 0.2 × SSC for 30 min. The slides were then incubated to biotinylated mouse anti-digoxigenin antibody at 37.5°C for 1 h followed by washing in 1 × PBS for 20 min at room temperature, and then to strepavidin-peroxidase at 37.5°C for 20 min followed by washing in 1 × PBS for 15 min at room temperature. Subsequently, the slides were developed with diaminobenzidine and then coun-

Table 1: Correlation of PSCA expression with Gleason score

	Intensity × frequency	
Gleason score	0–6 (%)	9 (%)
2–4	5 (83)	1 (17)
5–7	19 (79)	5 (21)
8–10	5 (28)	13 (72)

Table 2: Correlation of PSCA expression with clinical stage

	Intensity × frequency	
Tumor stage	0–6 (%)	9 (%)
≤B	27 (67.5)	13 (32.5)
≥C	2 (25)	6 (75)

terstained with hematoxylin to localize the hybridization signals. Sections hybridized with the sense control probes routinely did not show any specific hybridization signal above background. All slides were hybridized with PBS to substitute for the probes as a negative control.

Scoring methods

To determine the correlation between the results of PSCA immunostaining and mRNA *in situ* hybridization, the same scoring manners are taken in the present study for PSCA protein staining by IHC and PSCA mRNA staining by ISH. Each slide was read and scored by two independently experienced urological pathologists using Olympus BX-41 light microscopes. The evaluation was done in a blinded fashion. For each section, five areas of similar grade were analyzed semiquantitatively for the fraction of cells staining. Fifty percent of specimens were randomly chosen and rescored to determine the degree of interobserver and intraobserver concordance. There was greater than 95% intra- and interobserver agreement.

The intensity of PSCA expression evaluated microscopically was graded on a scale of 0 to 3+ with 3 being the highest expression observed (0, no staining; 1+, mildly intense; 2+, moderately intense; 3+, severely intense). The staining density was quantified as the percentage of cells staining positive for PSCA with the primary antibody or hybridization probe, as follows: 0 = no staining; 1 = positive staining in <25% of the sample; 2 = positive staining in 25%–50% of the sample; 3 = positive staining in >50% of the sample. Intensity score (0 to 3+) was multiplied by the density score (0–3) to give an overall score of 0–9 [1,5]. In this way, we were able to differentiate specimens that may have had focal areas of increased staining from those that had diffuse areas of increased staining [6]. The overall score for each specimen was then categorically assigned to one of the following groups: 0 score, negative expression; 1–2 scores, weak expression; 3–6 scores, moderate expression; 9 score, strong expression.

Statistical analysis

Intensity and density of PSCA protein and mRNA expression in BPH, PIN and Pca tissues were compared using the Chi-square and Student's *t*-test. Univariate associations between PSCA expression and Gleason score, clinical stage and progression to androgen-independence were calculated using Fisher's Exact Test. For all analyses, $p < 0.05$ was considered statistically significant.

Results

PSCA expression in BPH

In general, PSCA protein and mRNA were expressed weakly in individual samples of BPH. Some areas of prostate expressed weak levels (composite score 1–2), whereas other areas were completely negative (composite score 0). Four cases (20%) of BPH had moderate expression of PSCA protein and mRNA (composite score 4–6) by IHC and ISH. In 2/20 (10%) BPH specimens, PSCA mRNA expression was moderate (composite score 3–6), but PSCA protein expression was weak (composite score

2) in one and negative (composite score 0) in the other. PSCA expression was localized to the basal and secretory epithelial cells, and prostatic stroma was almost negative staining for PSCA protein and mRNA in all cases examined.

PSCA expression in PIN

In this study, we detected weak or negative expression of PSCA protein and mRNA (≤2 scores) in 7 of 9 (77.8%) low grade PIN and in 2 of 11 (18.2%) HGPIN, and moderate expression (3–6 scores) in the rest 2 low grade PIN and 5 of 11 (45.5%) HGPIN. One HGPIN with moderate PSCA mRNA expression (6 score) was found weak staining for PSCA protein (2 score) by IHC. Strong PSCA protein and mRNA expression (9 score) were detected in the remaining 3 of 11 (27.3%) HGPIN. There was a statistically significant difference of PSCA protein and mRNA expression levels observed between HGPIN and BPH ($p < 0.05$), but no statistical difference reached between low grade PIN and BPH ($p > 0.05$).

PSCA expression in Pca

In order to determine if PSCA protein and mRNA can be detected in prostate cancers and if PSCA expression levels are increased in malignant compared with benign glands, Forty-eight paraffin-embedded Pca specimens were analysed by IHC and ISH. It was shown that 19 of 48 (39.6%) Pca samples stained very strongly for PSCA protein and mRNA with a score of 9 and another 21 (43.8%) specimens displayed moderate staining with scores of 4–6 (Figure 1). In addition, 4 specimens with moderate to strong PSCA mRNA expression (scores of 4–9) had weak protein staining (a score of 2) by IHC analyses. Overall, Pca expressed a significantly higher level of PSCA protein and mRNA than any other specimen category in this study ($p < 0.05$, compared with BPH and PIN respectively). The result demonstrates that PSCA protein and mRNA are overexpressed by a majority of human Pca.

Correlation of PSCA expression with Gleason score in Pca

Using the semi-quantitative scoring method as described in Materials and Methods, we compared the expression level of PSCA protein and mRNA with Gleason grade of Pca, as shown in Table 1. Prostate adenocarcinomas were graded by Gleason score as 2–4 scores = well-differentiation, 5–7 scores = moderate-differentiation and 8–10 scores = poor-differentiation [7]. Seventy-two percent of Gleason scores 8–10 prostate cancers had very strong staining of PSCA compared to 21% with Gleason scores 5–7 and 17% with 2–4 respectively, demonstrating that poorly differentiated Pca had significantly stronger expression of PSCA protein and mRNA than moderately and well differentiated tumors ($p < 0.05$). As depicted in Figure 1, IHC and ISH analyses showed that PSCA protein and mRNA expression in several cases of poorly differentiated Pca were particularly prominent, with more intense and uniform staining. The results indicate that PSCA expression increases significantly with higher tumor grade in human Pca.

Correlation of PSCA expression with clinical stage in Pca

With regards to PSCA expression in every stage of Pca, we showed the results in Table 2. Seventy-five percent of locally advanced and node positive cancers (i.e. C-D stages) expressed statistically high levels of PSCA versus 32.5% that were organ confined (i.e. A-B stages) ($p < 0.05$). The data demonstrate that PSCA expression increases significantly with advanced tumor stage in human Pca.

Correlation of PSCA expression with androgen-independent progression of Pca

All 9 specimens of androgen-independent prostate cancers stained positive for PSCA protein and mRNA. Eight specimens were obtained from patients managed prior to androgen ablation therapy. Seven of eight (87.5%) of these androgen-independent prostate cancers were in the strongest staining category (score = 9), compared with three out of eight (37.5%) of patients with androgen-dependent cancers ($p < 0.05$). The results demonstrate that PSCA expression increases significantly with progression to androgen-independence of human Pca.

It is evident from the results above that within a majority of human prostate cancers the level of PSCA protein and mRNA expression correlates significantly with increasing grade, worsening stage and progression to androgen-independence.

Correlation of PSCA immunostaining and mRNA in situ hybridization

In all 88 specimens surveyed herein, we compared the results of PSCA IHC staining with mRNA ISH analysis. Positive staining areas and its intensity and density scores evaluated by IHC were identical to those seen by ISH in 79 of 88 (89.8%) specimens (18/20 BPH, 19/20 PIN and 42/48 Pca respectively). Importantly, 27/27 samples with PSCA mRNA composite scores of 0–2, 32/36 samples with scores of 3–6 and 22/24 samples with a score of 9 also had PSCA protein expression scores of 0–2, 3–6 and 9 respectively. However, in 5 samples with PSCA mRNA overall scores of 3–6 and in 2 with scores of 9 there were less or negative PSCA protein expression (i.e. scores of 0–4), suggesting that this may reflect posttranscriptional modification of PSCA or that the epitopes recognized by PSCA mAb may be obscured in some cancers. The data demonstrate that the results of PSCA immunostaining were consistent with those of mRNA ISH analysis, showing a high degree of correlation between PSCA protein and mRNA expression.

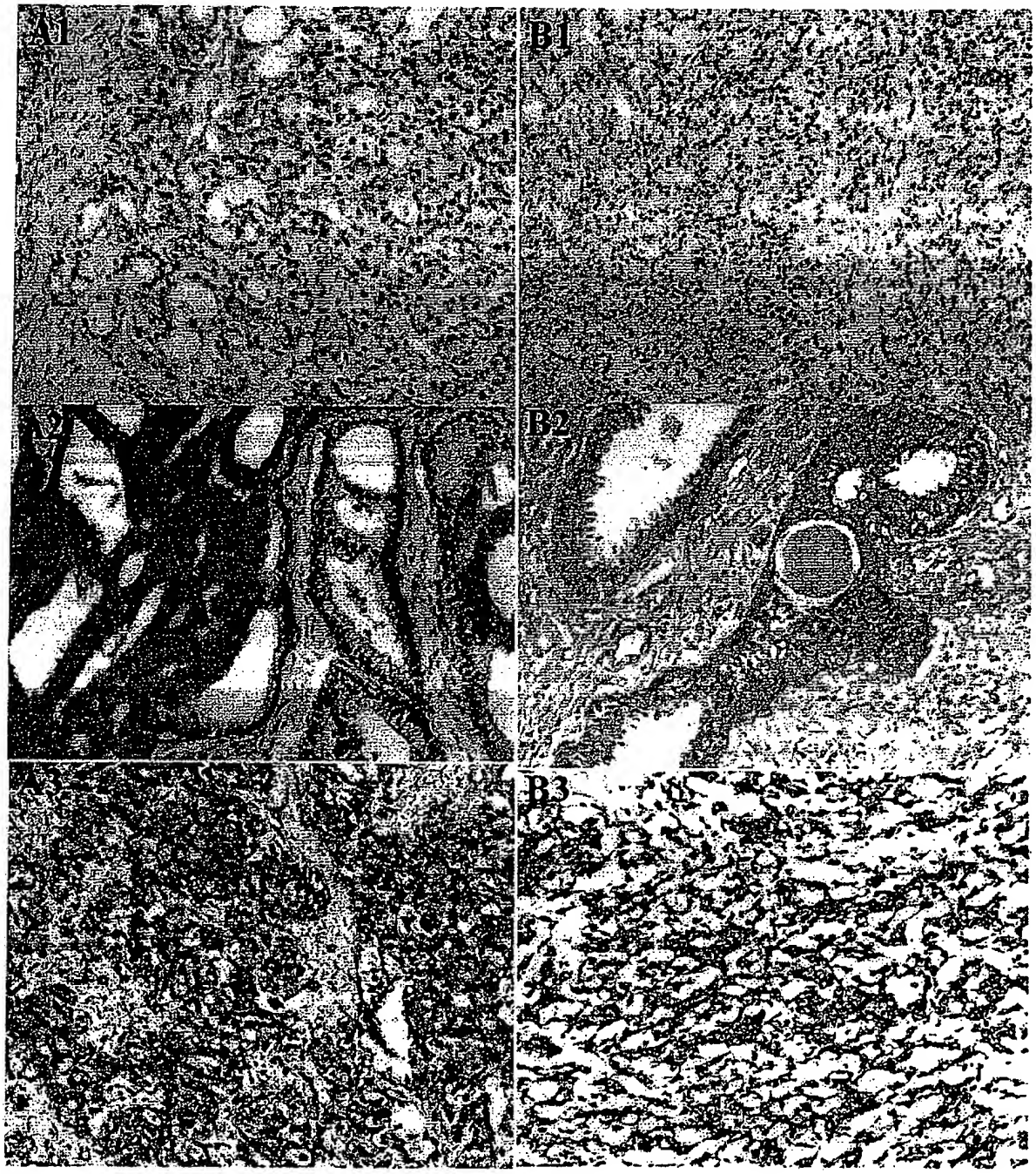

Figure 1

Representatives of PSCA IHC and ISH staining in Pca (A. IHC staining, B. ISH staining, ×200 magnification). A_1, B_1: negative control of IHC and ISH. PBS replacing the primary antibody (A_1) and hybridization with a sense PSCA probe (B_1) showed no background staining. A_2, B_2: a moderately differentiated Pca (Gleason score = 3+3 = 6) with moderate staining (composite score = 6) in all malignant cells; A_2: IHC shows not only cell surface but also apparent cytoplasmic staining of PSCA protein. A_3, B_3: a poorly differentiated Pca (Gleason score = 4+4 = 8) with very strong staining (composite score = 9) in all malignant cells.

Discussion

PSCA is homologous to a group of cell surface proteins that mark the earliest phase of hematopoietic development. PSCA mRNA expression is prostate-specific in normal male tissues and is highly up-regulated in both androgen-dependent and-independent Pca xenografts (LAPC-4 tumors). We hypothesize that PSCA may play a role in Pca tumorigenesis and progression, and may serve as a target for Pca diagnosis and treatment. In this study, IHC and ISH showed that in general there were weak or absent PSCA protein and mRNA expression in BPH and low grade PIN tissues. However, PSCA protein and mRNA are widely expressed in HGPIN, the putative precursor of invasive Pca, suggesting that up-regulation of PSCA is an early event in prostate carcinogenesis. Recently, Reiter RE et al [1], using ISH analysis, reported that 97 of 118 (82%) HGPIN specimens stained strongly positive for PSCA mRNA. A very similar finding was seen on mouse PSCA (mPSCA) expression in mouse HGPIN tissues by Tran C. P et al [8]. These data suggest that PSCA may be a new marker associated with transformation of prostate cells and tumorigenesis.

We observed that PSCA protein and mRNA are highly expressed in a large percentage of human prostate cancers, including advanced, poorly differentiated, androgen-independent and metastatic cases. Fluorescence-activated cell sorting and confocal/ immunofluorescent studies demonstrated cell surface expression of PSCA protein in Pca cells [9]. Our IHC expression analysis of PSCA shows not only cell surface but also apparent cytoplasmic staining of PSCA protein in Pca specimens (Figure 1). One possible explanation for this is that anti-PSCA antibody can recognize PSCA peptide precursors that reside in the cytoplasm. Also, it is possible that the positive staining that appears in the cytoplasm is actually from the overlying cell membrane [5]. These data seem to indicate that PSCA is a novel cell surface marker for human Pca.

Our results show that elevated level of PSCA expression correlates with high grade (i.e. poor differentiation), increased tumor stage and progression to androgen-independence of Pca. These findings support the original IHC analyses by Gu Z et al [9], who reported that PSCA protein expressed in 94% of primary Pca and the intensity of PSCA protein expression increased with tumor grade, stage and progression to androgen-independence. Our results also collaborate the recent work of Han KR et al [10], in which the significant association between high PSCA expression and adverse prognostic features such as high Gleason score, seminal vesicle invasion and capsular involvement in Pca was found. It is suggested that PSCA overexpression may be an adverse predictor for recurrence, clinical progression or survival of Pca. Hara H et al [11] used RT-PCR detection of PSA, PSMA and PSCA in 1 ml of peripheral blood to evaluate Pca patients with poor prognosis. The results showed that among 58 PCa patients, each PCR indicated the prognostic value in the hierarchy of PSCA>PSA>PSMA RT-PCR, and extraprostatic cases with positive PSCA PCR indicated lower disease-progression-free survival than those with negative PSCA PCR, demonstrating that PSCA can be used as a prognostic factor. Dubey P et al [12] reported that elevated numbers of PSCA + cells correlate positively with the onset and development of prostate carcinoma over a long time span in the prostates of the TRAMP and PTEN +/- models compared with its normal prostates. Taken together with our present findings, in which PSCA is overexpressed from HGPIN to almost frank carcinoma, it is reasonable and possible to use increased PSCA expression level or increased numbers of PSCA-positive cells in the prostate samples as a prognostic marker to predict the potential onset of this cancer. These data raise the possibility that PSCA may have diagnostic utility or clinical prognostic value in human Pca.

The cause of PSCA overexpression in Pca is not known. One possible mechanism is that it may result from PSCA gene amplification. In humans, PSCA is located on chromosome 8q24.2 [1], which is often amplified in metastatic and recurrent Pca and considered to indicate a poor prognosis [13-15]. Interestingly, PSCA is in close proximity to the c-myc oncogene, which is amplified in >20% of recurrent and metastatic prostate cancers [16,17]. Reiter RE et al [18] reported that PSCA and MYC gene copy numbers were co-amplified in 25% of tumors (five out of twenty), demonstrating that PSCA overexpression is associated with PSCA and MYC coamplification in Pca. Gu Z et al [9] recently reporteted that in 102 specimens available to compare the results of PSCA immunostaining with their previous mRNA ISH analysis, 92 (90.2%) had identically positive areas of PSCA protein and mRNA expression. Taken together with our findings, in which we detected moderate to strong expression of PSCA protein and mRNA in 34 of 40 (85%) Pca specimens examined simultaneously by IHC and ISH analyses, it is demonstrated that PSCA protein and mRNA overexpressed in human Pca, and that the increased protein level of PSCA was resulted from the upregulated transcription of its mRNA.

At present, the regulation mechanisms of human PSCA expression and its biological function are yet to be elucidated. PSCA expression may be regulated by multiple factors [18]. Watabe T et al [19] reported that transcriptional control is a major component regulating PSCA expression levels. In addition, induction of PSCA expression may be regulated or mediated through cell-cell contact and protein kinase C (PKC) [20]. Homologues of PSCA have diverse activities, and have themselves been involved in

carcinogenesis. Signalling through SCA-2 has been demonstrated to prevent apoptosis in immature thymocytes [21]. Thy-1 is involved in T cell activation and transducts signals through src-like tyrosine kinases [22]. Ly-6 genes have been implicated both in tumorigenesis and in cell-cell adhesion [23-25]. Cell-cell or cell-matrix interaction is critical for local tumor growth and spread to distal sites. From its restricted expression in basal cells of normal prostate and its homology to SCA-2, PSCA may play a role in stem/progenitor cell function, such as self-renewal (i.e. anti-apoptosis) and/or proliferation [1]. Taken together with the results in the present study, we speculate that PSCA may play a role in tumorigenesis and clinical progression of Pca through affecting cell transformation and proliferation. From our results, it is also suggested that PSCA as a new cell surface antigen may have a number of potential uses in the diagnosis, therapy and clinical prognosis of human Pca. PSCA overexpression in prostate biopsies could be used to identify patients at high risk to develop recurrent or metastatic disease, and to discriminate cancers from normal glands in prostatectomy samples. Similarly, the detection of PSCA-overexpressing cells in bone marrow or peripheral blood may identify and predict metastatic progression better than current assays, which identify only PSA-positive or PSMA-positive prostate cells.

In summary, we have shown in this study that PSCA protein and mRNA are maintained in expression from HGPIN through all stages of Pca in a majority of cases, which may be associated with prostate carcinogenesis and correlate positively with high tumor grade (poor cell differentiation), advanced stage and androgen-independent progression. PSCA protein overexpression is due to the upregulation of its mRNA transcription. The results suggest that PSCA may be a promising molecular marker for the clinical prognosis of human Pca and a valuable target for diagnosis and therapy of this tumor.

Competing interests

None declared.

References

1. Reiter RE, Gu Z, Watabe T, Thomas G, Szigeti K, David E, Wahl M, Nisitani S, Yamashiro J, Le Beau MM, Loda M, Witte ON: **Prostate stem cell antigen: a cell surface marker overexpressed in prostate cancer.** *Proc Natl Acad Sci USA* 1998, **95**:1735-1740.
2. Gleason DF: **Histologic grading and clinical staging of prostatic carcinoma.** In: *Urologic Pathology: The Prostate* Edited by: Tannebaum M. Philadelphia, Lea & Febiger; 1977:171-197.
3. Brawer MK: **Prostatic intraepithelial neoplasia: a premalignant lesion.** *Hum Pathol* 1992, **23**:242-248.
4. Amin MB, Ro JY, Ayala AC: **Prostatic intraepithelial neoplasia: relationship to adenocarcinoma of prostate.** *Pathol Annu* 1994, **29**:1-30.
5. Amara N, Palapattu GS, Schrage M, Gu Z, Thomas GV, Dorey F, Said J, Reiter RE: **Prostate stem cell antigen is overexpressed in human transitional cell carcinoma.** *Cancer Res* 2001, 61:4660-4665.
6. Hanas JS, Lerner MR, Lightfoot SA, Raczkowski C, Kastens DJ, Brackett DJ, Postier RG: **Expression of the cyclin-dependent kinase inhibitor p21 (WAF1/CIP1) and p53 tumor suppressor in dysplastic progression and adenocarcinoma in Barrett esophagus.** *Cancer (Phila)* 1999, **86**:756-763.
7. Egevad L, Gramfors T, Karlberg L: **Prognostic value of the Gleason score in prostate cancer.** *BJU Int* 2002, **89**:538-542.
8. Tran CP, Lin C, Yamashiro J, Reiter RE: **Prostate stem cell antigen is a marker of late intermediate prostate epithelial cells.** *Mol Cancer Res* 2002, **1**:113-121.
9. Gu Z, Thomas G, Yamashiro J, Shintaku IP, Dorey F, Raitano A, Witte ON, Said JW, Loda M, Reiter RE: **Prostate stem cell antigen (PSCA) expression increases with high Gleason score, advanced stage and bone metastasis in prostate cancer.** *Oncogene* 2000, **19**:1288-1296.
10. Han KR, Seligson DB, Liu X, Horvath S, Shintaku PI, Thomas GV, Said JW, Reiter RE: **Prostate stem cell antigen expression is associated with gleason score, seminal vesicle invasion and capsular invasion in prostate cancer.** *J Urol* 2004, **171**:1117-1121.
11. Hara H, Kasahara T, Kawasaki T, Bilim V, Obara K, Takahashi K, Tomita Y: **Reverse Transcription-Polymerase Chain Reaction Detection of Prostate-specific Antigen, Prostate-specific Membrane Antigen, and Prostate Stem Cell Antigen in One Milliliter of Peripheral Blood.** *Clin Cancer Res* 2002, **8**:1794-1799.
12. Dubey P, Wu H, Reiter RE, Witte ON: **Alternative pathways to prostate carcinoma activate prostate stem cell antigen expression.** *Cancer Res* 2001, **61**:3256-3261.
13. Visa korpi T, Kallioniemi AH, Syvanen AC, Hyytinen ER, Karhu R, Tammela T, Isola JJ, Kallioniemi OP: **Genetic changes in primary and recrrent prostate cancer by comparative genomic hybridization.** *Cancer Res* 1995, **55**:342-347.
14. Sato K, Qian J, Slezak JM, Lieber MM, Bostwick DG, Bergstralh EJ, Jenkins RB: **Clinical significance of alterations of chromosome 8 in high-grade, advanced, nonmetastatic prostate carcinoma.** *J Natl Cancer Inst* 1999, **91**:1574-1580.
15. Van Den Berg C, Guan XY, Von Hoff D, Jenkins R, Bittner , Griffin C, Kallioniemi O, Visakorpi , McGill , Herath J, Epstein J, Sarosdy M, Meltzer P, Trent J: **DNA sequence amplification in human prostate cancer identified by chromosome microdissection: potential prognostic implications.** *Clin Cancer Res* 1995, **1**:11-18.
16. Jenkins RB, Qian J, Lieber MM, Bostwick DG: **Detection of c-myc oncogene amplification and chromosomal anomalies in metastatic prostatic carcinoma by fluorescence in situ hybridization.** *Cancer Res* 1997, **57**:524-531.
17. Nupponen NN, Kakkola L, Koivisto P, Visakorpi T: **Genetic alterations in hormone-refractory recurrent prostate carcinomas.** *Am J Pathol* 1998, **153**:141-148.
18. Reiter RE, Sato I, Thomas G, Qian J, Gu Z, Watabe T, Loda M, Jenkins RB: **Coamplification of prostate stem cell antigen (PSCA) and MYC in locally advanced prostate cancer.** *Genes Chromosomes Cancer* 2000, **27**:95-103.
19. Watabe T, Lin M, Donjacour AA, Cunha GR, Witte ON, Reiter RE: **Growth, regeneration, and tumorigenesis of the prostate activates the PSCA promoter.** *Proc Natl Acad Sci USA* 2002, **99**:401-406.
20. Bahrenberg G, Brauers A, Joost HG, Jakse G: **PSCA expression is regulated by phorbol ester and cell adhesion in the bladder carcinoma cell line RT112.** *Cancer Lett* 2001, **168**:37-43.
21. Noda S, Kosugi A, Saitoh S, Narumiya S, Hamaoka T: **Protection from anti-TCR/CD3-induced apoptosis in immature thymocytes by a signal through thymic shared antigen-1/stem cell antigen-2.** *J Exp Med* 1996, **183**:2355-2360.
22. Thomas PM, Samelson LE: **The glycophosphatidylinositol-anchored Thy-1 molecule interacts with the p60fyn protein tyrosine kinase in T cells.** *J Biol Chem* 1992, **267**:12317-12322.
23. Bamezai A, Rock KL: **Overexpressed Ly-6A.2 mediated cell-cell adhesion by binding a ligand expressed on lymphoid cells.** *Proc Natl Acad Sci USA* 1995, **92**:4294-4298.
24. Katz BZ, Eshel R, Sagi-Assif O, Witz IP: **An association between high Ly-6A/E expression on tumor cells and a highly malignant phenotype.** *Int J Cancer* 1994, **59**:684-691.
25. Brakenhoff RH, Gerretsen M, Knippels EM, van Dijk M, van Essen H, Weghuis DO, Sinke RJ, Snow GB, van Dongen GA: **The human E48 antigen, highly homologous to the murine Ly-6 antigen ThB, is a GPI-anchored molecule apparently involved in keratinocyte cell-cell adhesion.** *J Cell Biol* 1995, **129**:1677-1689.

BMC Gastroenterology

BioMed Central

Research article

Open Access

Cyclin A and cyclin D1 as significant prognostic markers in colorectal cancer patients

Abeer A Bahnassy*[1], Abdel-Rahman N Zekri[2], Soumaya El-Houssini[1], Amal MR El-Shehaby[3], Moustafa Raafat Mahmoud[1], Samira Abdallah[4] and Mostafa El-Serafi[5]

Address: [1]Pathology Department, National Cancer Institute, Cairo University, Cairo, Egypt, [2]Virology and Immunology Unit, Cancer Biology Department, National Cancer Institute, Cairo University, Cairo, Egypt, [3]Biochemistry Department, Kasr El-Eini School of Medicine, Cairo University, Cairo, Egypt, [4]Pathologyy Department, Kasr El-Eini School of Medicine, Cairo University, Cairo, Egypt and [5]Medical Oncology Department, National Cancer Institute, Cairo University, Cairo, Egypt

Email: Abeer A Bahnassy* - chaya2000@hotmail.com; Abdel-Rahman N Zekri - ncizakri@starnet.com.eg; Soumaya El-Houssini - chaya2000@hotmail.com; Amal MR El-Shehaby - chaya2000@hotmail.com; Moustafa Raafat Mahmoud - ncizakri@starnet.com.eg; Samira Abdallah - chaya2000@hotmail.com; Mostafa El-Serafi - melserafi@starnet.com.eg

* Corresponding author

Published: 23 September 2004

BMC Gastroenterology 2004, 4:22 doi:10.1186/1471-230X-4-22

Received: 25 April 2004

Accepted: 23 September 2004

This article is available from: http://www.biomedcentral.com/1471-230X/4/22

Abstract

Background: Colorectal cancer is a common cancer all over the world. Aberrations in the cell cycle checkpoints have been shown to be of prognostic significance in colorectal cancer.

Methods: The expression of *cyclin D1*, *cyclin A*, *histone H3* and *Ki-67* was examined in 60 colorectal cancer cases for co-regulation and impact on overall survival using immunohistochemistry, southern blot and in situ hybridization techniques. Immunoreactivity was evaluated semi quantitatively by determining the staining index of the studied proteins.

Results: There was a significant correlation between *cyclin D1* gene amplification and protein overexpression (concordance = 63.6%) and between *Ki-67* and the other studied proteins. The staining index for *Ki-67*, *cyclin A and D1* was higher in large, poorly differentiated tumors. The staining index of *cyclin D1* was significantly higher in cases with deeply invasive tumors and nodal metastasis. Overexpression of *cyclin A* and *D1* and amplification of *cyclin D1* were associated with reduced overall survival. Multivariate analysis shows that *cyclin D1 and A* are two independent prognostic factors in colorectal cancer patients.

Conclusions: Loss of cell cycle checkpoints control is common in colorectal cancer. *Cyclin A and D1* are superior independent indicators of poor prognosis in colorectal cancer patients. Therefore, they may help in predicting the clinical outcome of those patients on an individual basis and could be considered important therapeutic targets.

Background

Colorectal cancer (CRC) is the third most common cancer in Western countries [1]. In Egypt, CRC has unique characteristics that differ from that reported in other countries of the western society. It was estimated that 35.6% of the Egyptian CRC cases are below 40 years of age and patients

usually present with advanced stage, high grade tumors that carry more mutations [2]. This uniquely high proportion of early-onset CRC, the early and continuous exposure to hazardous environmental agents, the different mutational spectrum and the prevalent consanguinity in Egypt justify further studies [3]. It was proved that most cancers result from accumulation of genetic alterations involving certain groups of genes, the majority of which are cell cycle regulators that either stimulate or inhibit cell cycle progression [1]. Cell proliferation allows orderly progression through the cell cycle, which is governed by a number of proteins including *cyclins* and *cyclin* dependent kinases [4,5]. The *cyclins* belong to a superfamily of genes whose products complex with various *cyclin*-dependent kinases (*cdks*) to regulate transitions through key checkpoints of the cell cycle [6]. Abnormalities of several *cyclins* have been reported in different tumor types, implicating, in particular, *cyclin A*, *cyclin E* and *cyclin D* [6,7].

Cyclin D1 is a G1 *cyclin* that regulates the transition from G1 to S phase since its peak level and maximum activity are reached during the G1 phase of the cell cycle. Whereas *cyclin A* is regarded a regulator of the transition to mitosis since it reaches its maximum level during the S and G2 phases [8]. The mechanisms likely to activate the oncogenic properties of the *cyclins* include chromosomal translocations, gene amplification and aberrant protein overexpression [7,9].

Several studies have shown that, *histone H3* mRNA expression can be used to identify the S phase fraction (SPF) through the in situ hybridization (ISH) technique [10,11]. The level of *histone H3* mRNA reaches its peak during the S phase and then drops rapidly at the G2 phase [12].

In face of the increasing incidence of CRC and its peculiar pattern in the Egyptian population, the present study was conducted to assess the role of *Ki-67* (pan-cell cycle marker), *cyclin D1* (G1 phase marker), *histone H3* mRNA (S phase marker), *cyclin A* (S to G2 phase marker) in CRC. The expression level of these markers was correlated to the clinicopathologic features and the overall survival of patients.

Table 1: Clinicopathological features of patients in relation to the staining index (SI) of *Ki-67*, *cyclin D1*, *cyclin A*, *histone H3*

		SI (mean + SD)			
Variables	*No. of cases*	*Ki-67*	*Cyclin D1*	*Cyclin A*	*Histone H3*
Sex					
Male	36	18.0 ± 6.4	6.7 ± 4.3	12.7 ± 5.7	10.7 ± 5.3
Female	24	20.1 ± 5.8	8.8 ± 8.4	10.0 ± 6.0	10.7 ± 5.4
Age (years)					
≥50	41	11.7 ± 6.0*	5.6 ± 5.2	10.0 ± 5.3	6.0 ± 5.0*
<50	19	23.8 ± 5.6	7.7 ± 6.8	13.6 ± 5.7	22.0 ± 5.2
Tumor size (cm)					
<5.0	33	12.2 ± 6.3*	5.3 ± 3.8*	11.5 ± 6.1*	10.3 ± 4.9*
≥5.0	27	30.1 ± 6.2	22.8 ± 7.2	28.6 ± 5.6	24.0 ± 5.6
Histology					
Normal	20	3.5 ± 2.0*	0.6 ± 0.2*	2.3 ± 1.1*	2.2 ± 0.9
Carcinoma	60	30.3 ± 6.2	24.9 ± 6.3	27.2 ± 5.8	10.7 ± 5.3
GI	15	11.7 ± 6.2	6.6 ± 4.0	10.0 ± 5.4	11.4 ± 4.9
GII	21	11.8 ± 5.6	8.9 ± 3.6	12.3 ± 6.5	7.8 ± 5.4
GIII	24	30.0 ± 4.3	22.0 ± 8.1	27.0 ± 4.9	11.5 ± 5.4
Lymph node					
Negative	33	19.5 ± 7.0	5.4 ± 5.3*	11.9 ± 6.5	12.3 ± 5.5
Positive	27	21.3 ± 4.9	20.6 ± 6.9	12.5 ± 5.0	14.2 ± 5.0
Depth of invasion					
m, sm	17	20.7 ± 6.7	3.1 ± 3.1*	11.9 ± 7.2	10.4 ± 5.1
beyond sm	43	21.9 ± 6.2	12.4 ± 6.5	12.2 ± 5.6	10.7 ± 5.4
Stage					
I	6	20.6 ± 6.7	5.7 ± 6.9	24.2 ± 6.9	11.1 ± 5.3
II	27	20.8 ± 6.9	5.3 ± 4.3	24.6 ± 6.0	10.4 ± 5.7
III	12	22.0 ± 5.4	7.7 ± 6.0	27.1 ± 5.2	10.4 ± 4.9
IV	15	24.7 ± 6.1	11.3 ± 9.6	27.5 ± 5.5	12.3 ± 6.2

* p. value < 0.05 (significant)

Methods

Tissue samples

Paraffin-embedded tumor tissues were obtained from 60 CRC patients (47 colon and 13 rectal carcinomas) that were diagnosed and treated at the National Cancer Institute, Cairo, Egypt during the period from January, 1997 to June, 2002. Clinicopathological data of the studied cases are illustrated in table 1. None of the patients received any chemotherapy or irradiation prior to surgery. Histological diagnosis of all cases was done by 2 independent pathologists according to the WHO Histological Classification. Tumors were staged according to the TNM staging system [13]. The depth of tumor invasion was classified as invasion of the mucosa including muscularis mucosa (m), invasion of the submucosa (sm), or invasion beyond the submucosa [8]. Normal colonic tissues were obtained from autopsy specimens (n = 20) and were used as a control. The actual survival rate of the patients was calculated from the date of resection to the date of death.

Immunohistochemistry

Four micron sections of each normal and tumor specimen were cut onto positive-charged slides; air dried overnight, de-paraffinized in xylene, hydrated through a series of graded alcohol and washed in distilled water and 0.01 PBS (pH 7.4). Slides were then processed for IHC as described by Handa et al. [8]. using the following antibodies: *Ki-67* (MIB-1, Dako), *cyclin A* (6E6; Novocastra, Newcastle-Upon-Tyne, UK) and *cyclin D1* (DCS-6, Dako). A case of invasive breast cancer was used as a positive control for *Ki-67* and *cyclin A* whereas a case of mantle cell lymphoma was used as a control for *cyclin D1*. Negative controls were obtained by replacing the primary antibody by non-immunized rabbit or mouse serum.

Brown nuclear staining was regarded as a positive result for all studied markers. The proportion of positively-stained cells and the intensity of staining were scored in tumor and normal colorectal mucosal sections at medium power (×200). The degree of positive tumor staining (percentage of positive tumor cells in the examined section) was scored from 1–6 and the staining intensity was scored from 0–6 according to the pattern of staining in the examined section. Staining index (SI) was calculated by multiplying the cellularity and staining scores as described by King et al. [14].

In situ hybridization

All tumor samples and 5 normal controls were assessed for *histone H3* mRNA by ISH using the commercially available 550 base fluorescein-labeled DNA probe (Dako, Carpinteria, CA) as described by Nagao et al., 1996. This probe hybridizes to the whole mRNA transcript of the human *histoneH3* gene including the5' and 3' un-translated regions. Scoring of *histone H3* mRNA was performed as for immunohistochemistry, however, hybridization signals were detected in the cytoplasm.

Molecular detection of cyclin D1 gene amplification

High molecular weight DNA was extracted from paraffin-embedded tissues of the tumor and normal colorectal mucosal samples as previously described [15]. The proportion of neoplastic and normal cells was determined in each tumor sample by examining hematoxylin and eosin-stained slides obtained from the edge of the specimen used for DNA extraction. Tumor samples were evaluated for amplification of *cyclin D1* if more than 75% of the examined sections were formed of neoplastic cells. Accordingly, 50 cases were eligible for the analysis. Ten micrograms of the extracted DNA was digested with *Eco*R1. DNA from selected cases was also digested with *Bgl*II and *Hind*III. Samples were separated on 0.8% agarose gels and transferred to Hybond-N membranes (Amersham Int., Amersham, UK). The membranes were hybridized with 50% formamide, 5 × SSC, 5 × Denhardt's, 500 μg/ml denatured salmon sperm DNA, 10% dextran sulphate and 10^6 cpm/ml of ^{32}P-labeled *PRAD-1* probe for 24 h. Membranes were washed with 2 × SSC, 0.1% SDS at room temperature for 30 min followed by 2 × SSC, 0.1% SDS at 60°C for 30 min and 0.1 × SSC, 0.1% SDS at 60°C for 1 h. Filters were autoradiographed using an intensifying screen at -70°C for 24–72 h. After being stripped free of the *PRAD-1* probe, the same blots were hybridized with ^{32}P-labeled *B-actin* probe to normalize against possible variations in the loading or transfer of DNA. The autoradiograms were analyzed using a densitometer. Intensities of *PRAD-1/cyclin D1* were normalized to the *β-actin* control bands. The degree of amplification was calculated from these normalized values. Amplification was considered when the signal of the tumor band was ≥2-fold the value of the matched normal mucosa [16].

Statistical analysis

The Mann-Whitney non-parametric test was used to compare the SIs of pairs of subjects whereas the Kruskal-wallis was used for categorial data. Correlation between indices was performed using a simple linear regression test. The Kaplan-Meier method was used to create survival curves which were analyzed by the log-rank test. The impact of different variables on survival was determined using the Cox proportional hazards model. *p*. values less than 0.05 were considered significant.

Results

The results of IHC are illustrated in figures 1 and 2. In general, the staining index (SIs) of all studied markers was higher in carcinomas than in normal colonic mucosal samples ($p = 0.0001$). Normal colorectal mucosa revealed positive imunostaining for *Ki-67* in the lower half of the crypts only. A heterogeneous staining pattern was

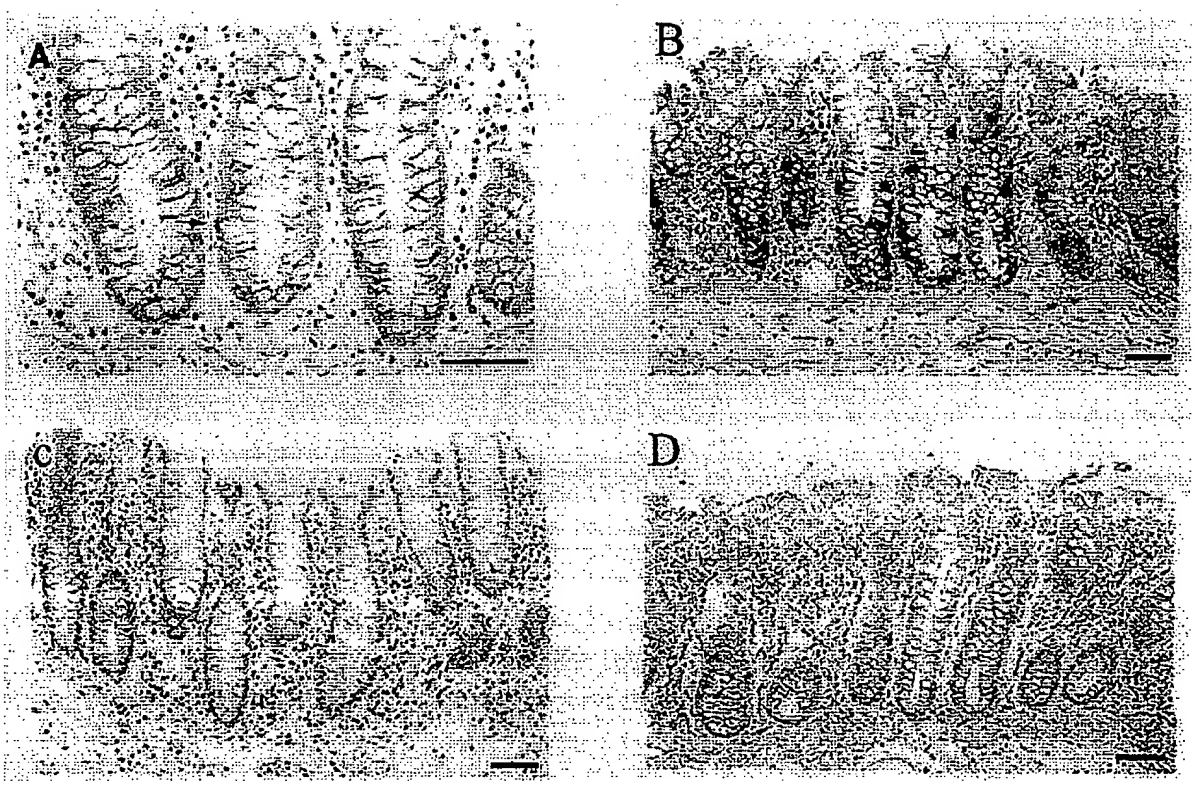

Figure 1
Normal colonic mucosa showing positive nuclear immunostaining for: (a) *cyclin D1*, (b) ISH of *histone H3* mRNA, (c) *Ki-67* and (d) *cyclin A*

detected in the neoplastic cells of well and moderately-differentiated adenocarcinomas whereas a diffuse homogeneous staining pattern was detected in poorly-differentiated carcinomas. The SI ranged from 10–40.2 (mean: 24.6 ± 6.5).

Immunostaining for *cyclin D1* was predominantly nuclear but cytoplasmic staining was detected in some cases. However, unless a nuclear staining was also detected, cases with cytoplasmic staining were considered negative. Normal colorectal mucosal samples were almost negative for *cyclin D1* whereas 41 out of the 60 (68.3%) CRC cases were positive. Marked heterogeneity was observed in well- and moderately-differentiated adenocarcinomas even within the same tumor. Poorly-differentiated carcinomas revealed a diffuse staining pattern with more darkly-stained nuclei. The SI ranged from 0.5–28.6 (mean: 9.3 ± 4.2).

Positive nuclear staining for *cyclin A* was detected in 80% (48/60) of CRC cases and in all non-neoplastic control samples. Positively-stained nuclei were confined to the lower half of the crypts in normal colonic mucosa and diffusely-dispersed in carcinomas. The SI ranged from 3.3–30.2 (mean: 15.1 ± 6.6).

Histone H3 mRNA was intensely expressed in the cytoplasm of all examined samples either neoplastic or non-neoplastic. The distribution of *histone H3* mRNA was similar to that of *cyclin A* and *Ki-67* however, the proportion of *histone H3* mRNA positive cells was less than that of *Ki-67*. The SI ranged from 1.8–24.2 (mean: 12.4 ± 5.3).

The *PRAD-1* probe detected 3 *Eco*RI fragments of 4.0, 2.2 and 2.0 and 1 *Bgl*II fragment of 15 Kb. *PRAD-1/cyclin D1* gene amplification was detected in 22/50 (44%) cases analyzed. The degree of amplification was heterogeneous

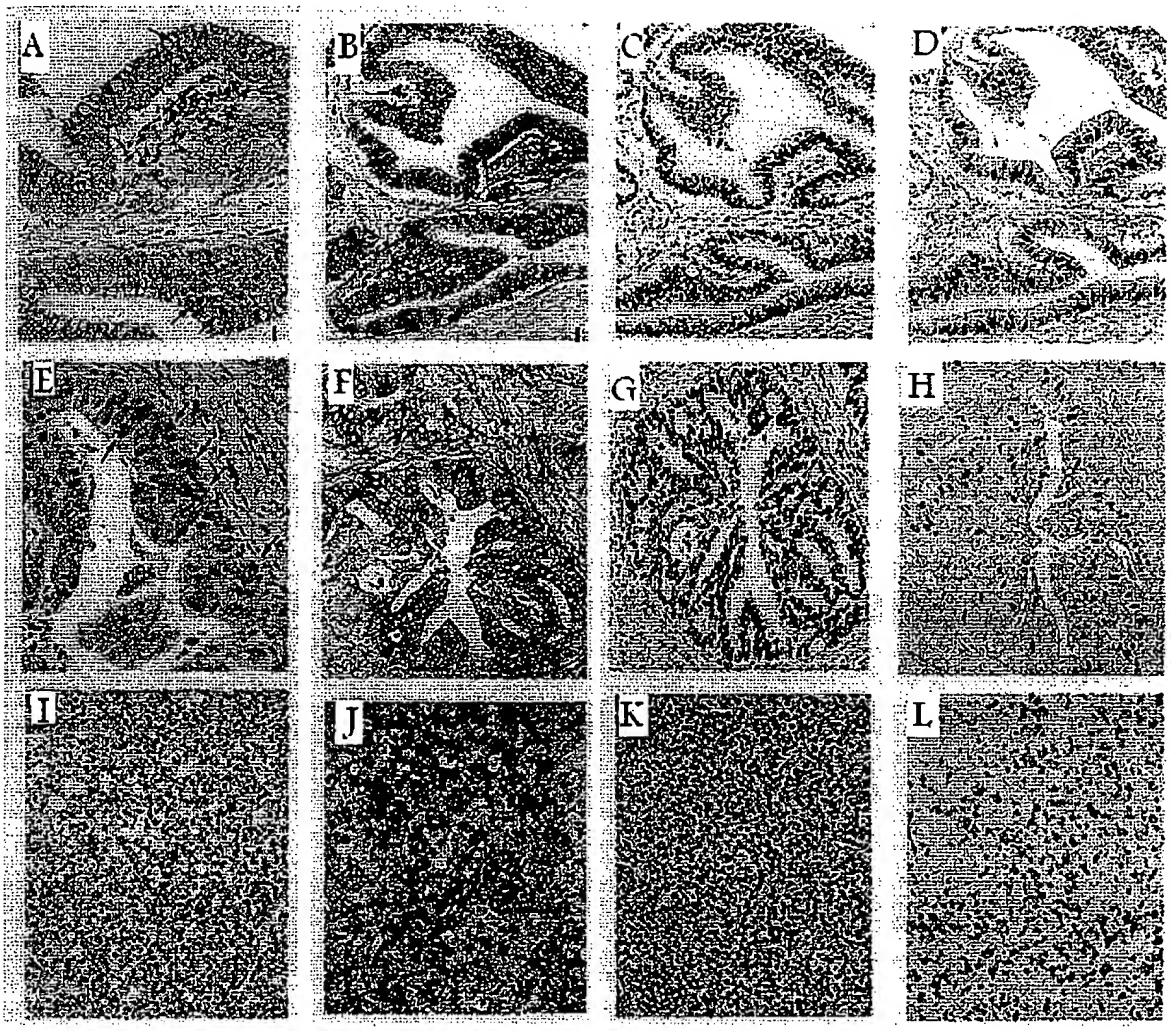

Figure 2
A case of well differentiated adenocarcinoma with positive immunostaining for: (a) *cyclin D1*, (b) *histone H3* mRNA, (c) *Ki-67*, and (d) *cyclin A*. Another case of moderately differentiated denocarcinoma with positive immunostaining for: (e) cyclin D1, (f) *histone H3* mRNA, (g) *Ki-67*, and (h) *cyclin A*. A case of poorly differentiated adenocarcinoma with diffuse staining for: (i) *cyclin D1*, (j) ISH of *histone H3* mRNA, (k) *Ki-67* and (l) *cyclin A*.

with 2–10 fold increase when compared to normal mucosal samples (Figure 3). Amplification was confirmed by other restriction enzymes.

Correlations

There was a significant correlation between *cyclin D1* gene amplification and protein overexpression. Out of the 22 cases that showed amplification 14 showed protein overexpression (concordance = 63.6%).

Linear regression analysis of SIs revealed a significant correlation between *Ki-67* and *cyclin D1*, *cyclin A*, *histone H3* as well as between the SIs of *cyclin A* and *histone H3* (p = 0.008, 0.0001, and 0.0001 respectively) (Figure 4). There was a significant relationship between the SI of both *Ki-67*

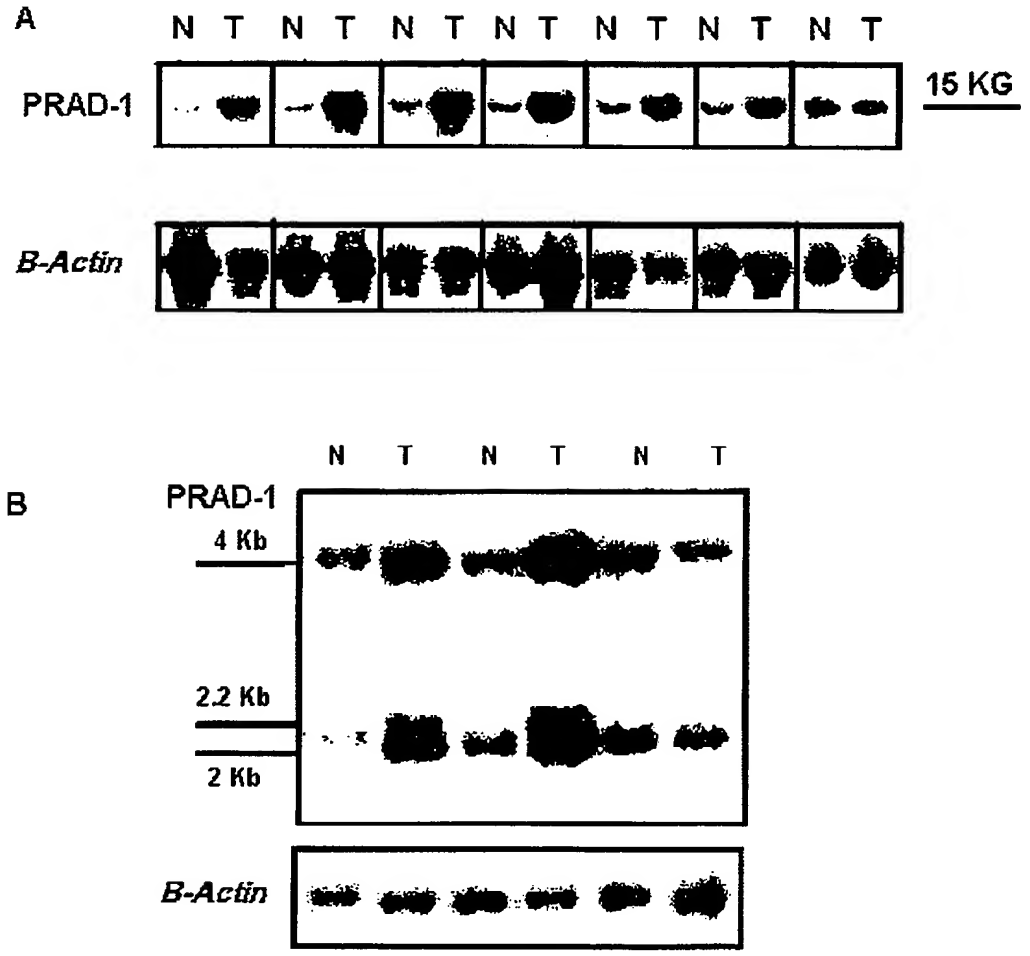

Figure 3

A: Southern blot analysis of normal mucosa (N) and their seven corresponding cases of colonic adenocarcinomas (T1–T7), cases No. 1, 2, 4, and 5 are poorly differentiated whereas cases No. 3, 6, and 7 are moderately differentiated. Genomic DNA was digested with *BglII*, fractionated by electrophoresis in agarose gel, transferred onto membranes and hybridized with *PRAD1* and *β-actin*. Tumors number 1–6 (Lanes 1–6) show different degrees of *PRAD1/cyclin D1* amplification, tumor number 7 (lane 7) was not amplified. **B:** Southern blot analysis of 3 cases of adenocarcinomas (T) and matched normal colonic mucosa (N). Genomic DNA was digested with *EcoRI*, fractionated by electrophoresis in agarose gel, transferred onto membranes and hybridized with *PRAD1* and *β-actin* probes for loading control. The identification of the 3 tumors is the same as in Fig. 3A with amplification of *PRAD1/cyclin D1* in tumors number 4, 5 (Lanes 1, 2) but not 7 (Lane 3).

and *cyclin A* and the degree of differentiation of tumors as well as the size of the tumor ($p < 0.001$ and $p < 0.01$ respectively). In addition, SI of *Ki-67* and *histone H3* were higher in patients <50 years than in those ≥50 years ($p < 0.05$) (table 1).

In addition table 2 shows a significant relationship between high *cyclin D1* SI and large, poorly-differentiated tumors, carcinomas with positive lymph node metastasis and deeply-invasive carcinomas ($p < 0.05$, $p < 0.001$, $p < 0.05$ and $p < 0.05$ respectively). Whereas *cyclin D1* gene amplification was significantly associated with an advanced disease stage since amplification was detected in

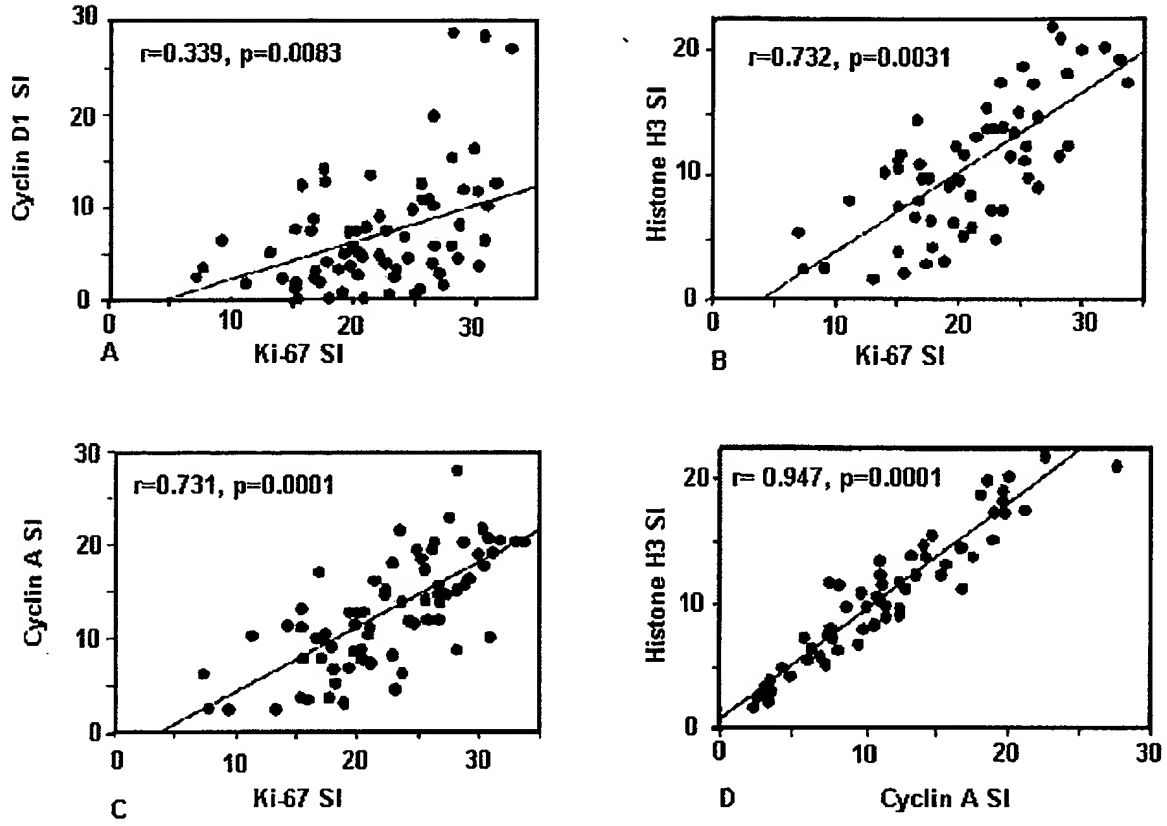

Figure 4
Correlation between the staining intensity of (a) *Ki-67 vs. cyclin D1*, (b) *Ki-67 vs. histone H3*, (c) *Ki-67 vs. cyclin A* and (d) *cyclin A vs. histone H3* mRNA expression.

10/15 (66.7%) of stage IV tumors compared to 12/45 (26.7%) of stage I-III tumors ($p = 0.002$). Similarly, DNA amplification was detected in 60.5% (26/43) of the carcinomas with extensive local invasion (beyond sm) but only in 23.5% (4/17) of the carcinomas with limited invasion (m, sm) ($p = 0.001$). A significant correlation was also present between *cyclin D1* gene amplification and the presence of lymph node metastasis ($p = 0.008$) as well as between the SI of *histone H3*, the size of the tumor and the patient's age ($p < 0.05$, $p < 0.001$ respectively). The SI was higher in tumors >5 cm in diameter and in patients <50 years.

Survival analysis

The mean follow-up period for all patients was 30 months (range: 1–66 months). Eighteen of 60 patients had already died by the time the study was completed. We defined the cutoff level for overexpression of each cell cycle marker at the point that showed the maximum difference of survival rate between the 2 groups separated by that point. Cox regression analysis revealed that *cyclin A* overexpression (our definition: SI ≥ 10.5), *cyclin D1* overexpression (our definition: SI ≥ 6.1), poorly differentiated histology, lymph node metastasis, TNM stage, tumor size and depth of invasion were all significant prognostic variables for survival (Table 3). The Kaplan-Meier survival curves for the subgroups of patients who are subdivided according to each marker's status are shown in Figure 5. Patient with tumors that showed *Ki-67* overexpression (our definition: SI ≥ 11.5) and *histone H3* overexpression (our definition: SI ≥ 8.2) tended to have poor prognosis but this did not reach a statistically significant level, however the overall survival was significantly lower in patient with *cyclin A* and *cyclin D1* overexpression. Cox multivari-

Table 2: The relation between *cyclin D1* overexpression vs *cyclin D1* amplification and clinicopathological prognostic markers.

Variables	*No. of cases*	*Cyclin D1 overexpression*	*Cyclin D1 Amplification*
Tumor size (cm)			
<5.0	33	5.3 ± 3.8*	13/33
≥5.0	27	22.8 ± 7.2 p *<0.05*	9/27 p *<0.236*
Histology			
GI	15	6.6 ± 4.0	7/15
GII	21	8.9 ± 3.6	8/21
GIII	24	22.0 ± 8.1 p *<0.001*	7/24 p *<0.075*
Lymph node			
Negative	33	5.4 ± 5.3*	6/33 (18.2%)
Positive	27	20.6 ± 6.9 p *<0.05*	16/27 (59.3%) p *<0.008*
Depth of invasion			
m, sm	17	3.1 ± 3.1*	4/17 (23.5%)
beyond sm	43	12.4 ± 6.5 p *<0.05*	26/43 (60.5%) p *<0.001*
Stage			
early	45	5.5 ± 10.1	12/45 (26.7%)
late	15	11.3 ± 9.6 P = *0.175*	10/15 (66.7%) p *<0.002*

Table 3: Uunivariate analysis of the relationship between survival and the tested markers

PredictiveVariables	*Median Survival*	*HR*	*CI*	*P*
Ki-67				
<11.5	36			
≥11.5	32	1.826	0.636 – 5.243	0.26
Cyclin D1				
<6.1	35			
≥6.1	18	7.246	1.007 – 45.150	0.03*
Histone H3				
<8.2	35			
≥8.2	29	4.639	0.854 – 25.196	0.07
Cyclin A				
<10.5	35			
≥10.5	15	7.820	1.017 – 60.122	0.02*
Histological grade				
Low	38			
High	10	7.331	2.696 – 19.940	0.0001*
Lymph node				
Negative	38			
Positive	15	6.826	1.973 – 23.621	0.002*
Stage				
I, II, III	38			
IV	12	6.378	1.842 – 22.083	0.001*
Tumor size (cm)				
<5.0	35			
≥5.0	13	4.835	1.386 – 16.868	0.01*
Depth of invasion				
T1, T2	36			
T3, T4	20	7.759	1.024 – 58.789	0.04*
Age (years)				
<50	38			
≥50	28	2.802	0.988 – 7.943	0.0526
Sex				
Male	38			
Female	36	0.696	00.274 – 1.766	0.4449

* p. value < 0.05 (significant)
HR: Hazard Ratio
CI: 95% confidence Interval

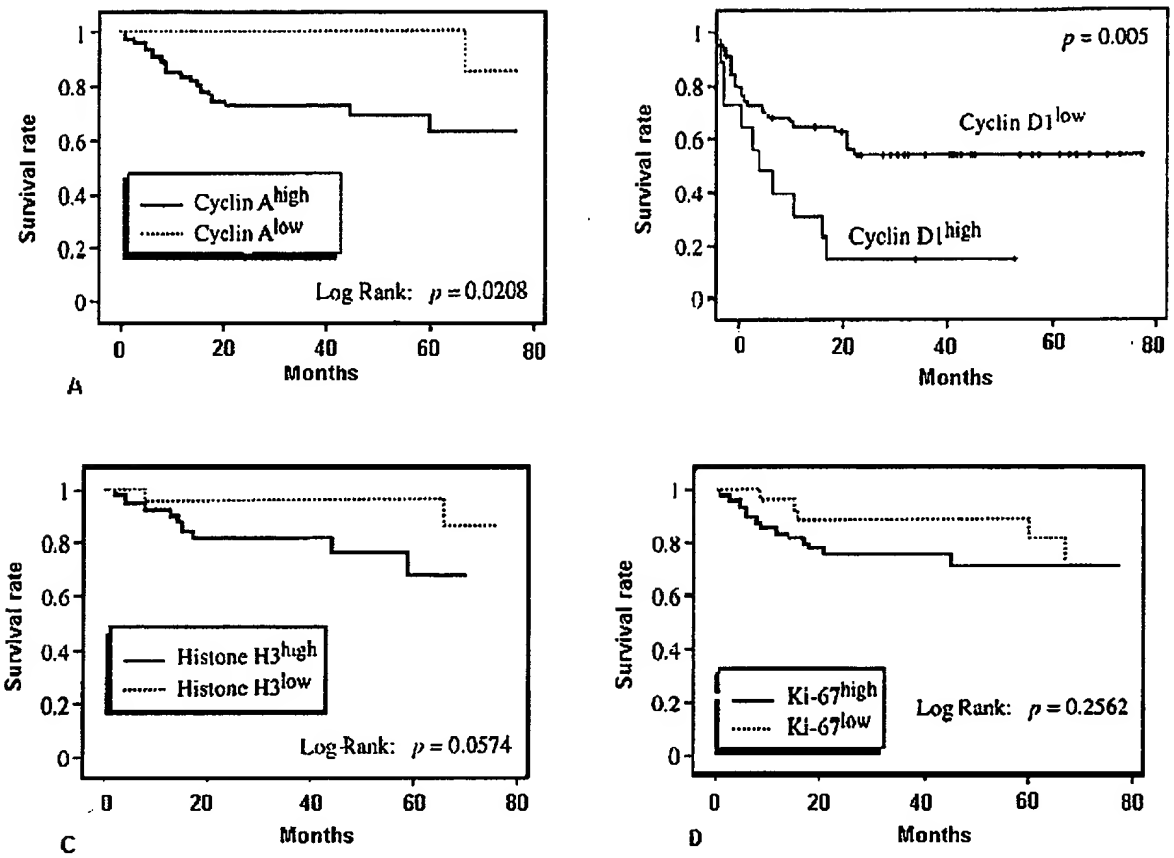

Figure 5
Kaplan-Meier survival curves for colorectal carcinoma. Overall survival is significantly lower in patients with (a) *cyclin A* and (b) *cyclin D1* overexpression. Patients with high SI for *histone H3* mRNA have poorer prognosis but this was not statistically significant (c). No significant difference was present between patients with high *Ki-67* SI and those with low *Ki-67* SI (d).

ate regression analysis revealed that lymph node metastasis, *cyclin A* and *cyclin D1* overexpression were independent negative prognostic factors after adjustment for the depth of tumor invasion, age and sex of the patient (Table 4).

Discussion

The proliferative activity of CRC cells has been investigated in several studies either by immunohistochemical determination of cell proliferation index using antibodies to some types of *cyclins* or by flowcytometric determination of the SPF of the cell cycle [8]. Although Leach et al. [17] did not find *cyclin D1* gene amplification in a panel of 47 CRC cell lines; its protein was overexpressed in about 30% of CRC cases that were included in the studies of Bartakova et al. [6] and Arber et al. [18]. In the former study [6]*cyclin D1* was aberrantly accumulated in a significant subset of human CRC cases and the cell lines derived from these cases were dependent on *cyclin* in their cell cycle progression. In the second study [18], overexpression of *cyclin D1* was detected in 30% of adenomatous polyps indicating that overexpression is a relatively early event in colon carcinogenesis which is possibly responsible for the pathological changes in the mucosa preceding neoplastic transformation. More recently, Holland et al. [19], Pasz-Walczak et al. [20] and Utsunomiya et al. [21] reported up-regulation of *cyclin D1* in 58.7%, 100% and 43% of their studied cases respectively.

Table 4: Multivariate analysis ofthe relationship between survival and thetested markers

PredictiveVariables	HR	CI	P
Cyclin D1	10.864	1.055 – 86.250	0.03*
(baseline < 6.1)	-	-	-
Cyclin A	13.886	1.012 – 190.579	0.0490*
(baseline < 10.5)	-	-	-
Positive Lymph node metastasis	3.921	1.057 – 14.472	0.0410*
Stage IV	3.411	1.048 – 12.083	0.03*
Depth of Invasion			
T3, T4	5.408	0.449 – 65.080	0.1836
Age (years)			
≥50	1.996	0.678 – 5.878	0.2310
Sex	0.910	0.315 – 2.358	0.8453

p. value < 0.05 (significant)
HR: Hazard Ratio
CI: 95% confidence Interval

In the present study, up-regulation of *cyclin D1* was detected in 68.3% of the cases. The SI was significantly higher in carcinomas than in normal colorectal mucosa and in poorly-differentiated adenocarcinomas it was approximately twice that of other histological types. Amplification and/or overexpression of *cyclin D1* significantly correlated with deeply invasive tumors and positive lymph node metastasis. Our results in this regards are consistent with previous studies [8,22]. In 2001, Holland et al. [19]. demonstrated that deregulation of *cyclin D1* and $p21^{waf}$ proteins are important in colorectal tumorigenesis and have implications for patient prognosis. Similarly McKay et al. [23] found that *cyclin D1* was the only protein in their panel (*cyclin D1, p53, p16, Rb-1, PCNA and p27*) that correlated with improved outcome in CRC patients. However, few studies failed to detect any correlation between *cyclin D1* overexpression and the clinicopathological factors in CRC [6,18]. This controversy in results could partially be explained by the difference in the sampling of studied cases. The present study included 24 cases of poorly differentiated adenocarcinoma, which is not common in other studies of CRC in western countries. This was possible because the majority of CRC cases diagnosed in Egypt are of high histological grade [3]. The correlation between *cyclin D1* overexpression and the high histological grade was also reported in other tumor types including non-small cell lung carcinomas [24] and squamous cell carcinomas of the larynx [16]. Another possible explanation for the observed controversy in the results of different studies is the detection method used.

In the present work, overexpression of *cyclin D1* was more common than gene amplification of the *PRAD-1/cyclin D1* gene with a 63.6% concordance. This was similarly reported by Bartakova et al. [6] who mentioned that there is a subset of CRC cases in which *cyclin D1* is overexpressed without *PRAD-1/cyclin D1* gene amplification. Consistent with this hypothesis are reports of elevated *cyclin D1* mRNA levels and immunohistochemically detectable accumulation of the protein in over one third of breast cancer cases at a frequency significantly higher than that deduced from DNA amplification studies [9,25]. These data imply that mechanisms other than gene amplification can also lead to deregulation and accumulation of *cyclin D1* in solid tumors.

So far, several studies were done to reveal the prognostic significance of *cyclin D1* overexpression in various carcinomas, including CRC [22]. However, these studies yielded conflicting results which could be attributed to organ heterogeneity. In our study, patients with tumors that exhibited *cyclin D1* overexpression tended to have poor prognosis.

It was reported that, patients with *cyclin A* positive carcinomas had significantly shorter median survival times. Handa et al. [8] were able to detect *cyclin A* overexpression in 77% of their CRC cases. They also demonstrated that, *cylcin A* could be used as a prognostic factor of CRC. More recently, Habermann et al. [26] studied cases of ulcerative colitis with and without an associated adenocarcinoma for the presence of *cyclin A* overexpression. They found that, *cyclin A* overexpression was higher in cases of ulcerative colitis with adenocarcinomas than in those without adenocarcinomas. Consequently, they concluded that, *cyclin A* could be used for monitoring ulcerative colitis patients and for the early detection of an emerging carcinoma in this high risk group of patients.

In our study, *cyclin A* was detected in 80% of the patients and Cox regression analysis showed that it could be used as a prognostic marker in CRC in addition to *cyclin D1*.

It would have been useful if we assessed the expression level of *cyclin A* by another technique (DNA amplification). This would have added more information regarding the gene status on one hand and confirmed the results of IHC on the other hand. Unfortunately, this was not possible because in most of the cases included in the present work, the extracted DNA was not sufficient to study *cyclin amplification* after the assessment of *cyclin D1*.

In 1996, Nagao et al. [11] reported that *histone H3* labeling index significantly correlated with ki-67 immunostaining and was high in poorly differentiated human hepatocellular carcinoma. This was similarly reported in the present work since we found a significant correlation between the SI of *histone H3* and *Ki-67*. However, no

statistically significant correlation was found between *histone H3* SI and any of the studied clinicopathological factors.

Although *Ki-67* immunostaining reflects the proliferative activity of CRC, it has not been recognized as a significant prognostic factor in this type of tumors [27,28]. However, Suzuki at al. [29] found a significant correlation between *Ki-67* labeling index and local invasion of CRC. In the present study there was a significant relationship between the SI of *Ki-67*, tumor size and grade. However, Kaplan-Meier survival curves showed no significant difference in survival rates between patients with- and without overexpression of *Ki-67*.

Conclusions

Our results demonstrate that *cyclin D1*, *cyclin A*, *histone H3* and *Ki-67* are overexpressed in a subset of CRC, however only *cyclin D1* and *cyclin A* overexpression correlates with poor differentiation and tumor progression. This indicates the superiority of *cyclin A* and *cyclin D1* as indicators of poor prognosis compared to *Ki-67* and *histone H3* mRNA in CRC. *Cyclin A and D1* could therefore be considered significant, independent prognostic factors in CRC patients. These findings are especially important in stage II patients since 25–30% of those patients have poor prognosis in spite of being node-negative. However, the standard clinicopathologic prognostic factors can not identify this subset accurately and therefore; there is a great demand for more accurate, individually-based, biological prognostic parameters that help in detecting this high risk group of patients who can benefit from an adjuvant therapy. If the findings of the present study are confirmed in a larger study, evaluation of *cyclin A and D1* may be applicable to clinical management of CRC, allowing the identification of patients with poor prognosis.

Competing interests

The author(s) declare that they have no competing interests.

List of abbreviations

CRC – Colorectal cancer

OS – overall survival

SI – staining index

SPF – S phase fraction

ISH – in situ hybridization

m – muscularis mucosa

sm – invasion of the sub mucosa

Authors' contributions

BA and ZA-R carried out the molecular genetic studies, designed, coordinated the study and drafted the manuscript. BA and El-HS carried out all the histopathological and immunohistochemical studies. El-SA participated in molecular genetic studies and drafted the manuscript. MM coordinated the study. El-SM carried out all the patient clinical data. All authors read and approved the final manuscript

References

1. Jiang GL, Huang S: **Adenovirus expressing RIZ1 in tumor suppressor gene therapy of microsatellite unstable colorectal cancers.** *Cancer Res* 2001, **61**:1796-1798.
2. Soliman AS, Bondy ML, Levin B, Hamza MR, Ismail K, Ismail S, Hammam HM, El-Hattab O, Kamal SM, Soliman AG, Dorgham LA, McPherson RS, Beasley RP: **Colorectal cancer in Egyptian patients under 40 years of age.** *Int J Cancer* 1997, **71**:26-30.
3. Soliman AS, Bondy ML, Guan Y, El-Badawy S, Mokhtar N, Bayomi S, Raouf AA, Ismail S, McPherson RS, Abdel-Hakim TF, Beasley PR, Levin B, Wei Q: **Reduced expression of mismatch repair genes in colorectal cancer patients in Egypt.** *Int J Oncol* 1998, **12**:1315-1319.
4. Cordon-Cardo C: **Mutations of cell cycle regulators. Biological and clinical implications for human neoplasia.** *Am J pathol* 1995, **147**:545-560.
5. Hunter T, Pines J: **Cyclins and cancer. II. Cyclin D and CDK inhibitors come of age.** *Cell* 1994, **79**:573-528.
6. Bartkova J, Lukas J, Strauss M, Bartek J: **The *PRAD-1/cyclin D1* oncogene product accumulates aberrantly in a subset of CRCs.** *Int J Cancer* 1994, **58**:568-573.
7. Motokura T, Arnold A: **Cyclins and oncogenesis.** *Biochim Biophys Acta* 1993, **1155**:63-78.
8. Handa K, Yamakawa M, Takeda H, Kimura S, Takahashi T: **Expression of the cell cycle markers in colorectal carcinoma: Superiority of cyclin A as an indicator of poor prognosis.** *Int J cancer* 1999, **84**:225-233.
9. Gillett C, Fantl V, Smith R, Fisher C, Bartek J, Dickson C, Barnes D, Peters G: **Amplification and overexpression of cyclin D1 in breast cancer detected by immunohistochemical staining.** *Cancer Res* 1994, **54**:1812-1817.
10. Gown AM, Jiang JJ, Matles H, Skelly M, Goodpaster T, Cass L, Reshatof M, Spaulding D, Coltrera DM: **Validation of the S-phase specificity of histone (H3) in situ hybridization in normal and malignant cells.** *J Histochem Cytochem* 1996, **44**:221-226.
11. Nagao T, Ishida Y, Kondo Y: **Determination of S-phase cells by in situ hybridization for histone H3 mRNA in hepatocellular carcinoma: correlation with histological grade and other cell proliferative markers.** *Mod Pathol* 1996, **9**:99-104.
12. Chou MY, Chang AL, McBride J, Donoff B, Gallagher GT, Wong DT: **A rapid method to determine proliferation patterns of normal and malignant tissues by H3 mRNA in situ hybridization.** *Am J Pathol* 1990, **136**:729-733.
13. Sobin LH, Wittekind C: **TNM classification of malignant tumors.** 5th edition. John Wiley, New York; 1997.
14. King RJ, Coffer AI, Gilbert J, Lewis K, Nash R, Millis R, Raju S, Taylor RW: **Histochemical studies with a monoclonal antibody raised against a partially purified soluble estradiol receptor preparation from human myometrium.** *Cancer Res* 1985, **45**:5728-5733.
15. Slebos RJ, Boerrigter L, Evers SG, Wisman P, Mooi WJ, Rodenhuis S: **A rapid and simple procedure for the routine detection of ras point mutations in formalin-fixed, paraffin-embedded tissues.** *Diag Moln Path* 1992, **1**:136-141.
16. Jares P, Fernandez P, Campo E, Nadal A, Bosch F, Aiza G, Nayach I, Traserra J, Cardesa A: ***PRAD-1/cyclin D1* gene amplification correlates with messenger RNA overexpression and tumor progression in human laryngeal carcinomas.** *Cancer Res* 1994, **54**:4813-4817.
17. Leach FS, Elledge SJ, Sherr CJ, Willson JK, Markowitz S, Kinzler KW, Vogelstein B: **Amplification of cyclin genes in colorectal carcinomas.** *Cancer Res* 1993, **53**:1986-1989.

18. Arber N, Hibshoosh H, Moss SF, Sutter T, Zhang Y, Begg M, Wang S, Weinstein IB, Holt PR: **Increased expression of cyclin D1 is an early event in multistage colorectal carcinogenesis.** *Gastroenterology* 1996, **110**:669-674.
19. Holland TA, Elder J, McCloud JM, Hall C, Deakin M, Fryer AA, Elder JB, Hoban PR: **Subcellular localization of cyclin D1 protein in colorectal tumors is associated with p21 (WAF1/CIP1) expression and correlates with patient survival.** *Int J Cancer* 2001, **95(5)**:302-306.
20. Pasz-Walczak G, Kordek R, faflik M: **P21(WAF1) expression in colorectal cancer: correlation with p53 and cyclin D1 expression, clinicopathological parameters and prognosis.** *Pathol Res Pract* 2001, **197(10)**:683-689.
21. Utsunomiya T, Doki Y, Takemoto H, Shiozaki H, Yano M, Sekimoto M, Tamura S, Yasuda T, Fujiwara Y, Monden M: **Correlation of beta-catenin and cyclin D1 expression in colon cancers.** *Oncology* 2001, **61(3)**:226-233.
22. Maeda K, Chung YS, Kang SM, Ogawa M, Onoda N, Nakata B, Nishiguchi Y, Ikehara T, Okuno M, Sowa M: **Overexpression of cyclin D1 and p53 is associated with disease recurrence in colorectal adenocarcinoma.** *Int JCancer* 1997, **74**:310-315.
23. McKay JA, Douglas JJ, Ross VG, Curran S, Loane JF, Ahmed FY, Cassidy J, McLeod HL, Murray GI: **Analysis of key cell cycle checkpoint proteins in colorectal tumors.** *J Pathol* 2002, **196**:386-393.
24. Mate JL, Ariza A, Aracil C, Lopez D, Isamat M, Perez-Piteira J, Navas-Palacios JJ: **Cyclin D1 overexpression in non-small cell lung carcinoma: correlation with Ki-67 labeling index and poor cytoplasmic differentiation.** *J Pathol* 1996, **180**:395-399.
25. Buckley MF, Sweeney KJ, Hamilton JA, Sini RL, Manning DL, Nicholson RI, DeFazio A, Watts CK, Musgrove EA, Sutherland RL: **Expression and amplification of cyclin genes in human breast cancer.** *Oncogene* 1993, **8**:2127-2133.
26. Habermann J, Lenander C, Roblick UJ, Kruger S, Ludwig D, Alalya A, Freitag S, Dumbgen L, Bruch HP, Stange E, Salo S, Tryggvason K, Auer G, Schimmelpenning H: **Ulcerative colitis and colorectal carcinoma: DNA profile, laminin-5 gamma 2 chain and cyclin A expression as early markers for risk assessment.** *Scand J Gastroenterol* 2001, **36**:751-758.
27. Kubota Y, Petras RE, Easley KA, Bauer TW, Tubbe RR, Fazio VW: **Ki-67-determined growth fraction versus standard staging and grading parameters in colorectal carcinoma. A multivariate analysis.** *Cancer* 1992, **70**:2602-2609.
28. Shain AA, Ro JY, Brown RW, Ordonez NG, Cleary KR, El-Naggar AK, Wilson P, Ayala AG: **Assessment of Ki-67-derived tumor proliferative activity in colorectal adenocarcinomas.** *Mod Pathol* 1994, **7**:17-22.
29. Suzuki H, Matsumoto K, Terabe M: **KI-67 antibody labeling index in colorectal carcinoma.** *J clin Gastroenterol* 1992, **15**:317-320.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-230X/4/22/prepub